Weakly Supervised 3D Hand Pose Estimation via Biomechanical Constraints

Weakly Supervised 3D Hand Pose Estimation via Biomechanical Constraints

Adrian Spurr[1,2]*, Umar Iqbal[2], Pavlo Molchanov[2],
Otmar Hilliges[1], and Jan Kautz[2]

[1] Advanced Interactive Technologies, ETH Zurich, Switzerland
[2] NVIDIA, Santa Clara, USA
{adrian.spurr, otmar.hilliges}@inf.ethz.ch
{uiqbal, pmolchanov, jkautz}@nvidia.com

Abstract. Estimating 3D hand pose from 2D images is a difficult, inverse problem due to the inherent scale and depth ambiguities. Current state-of-the-art methods train fully supervised deep neural networks with 3D ground-truth data. However, acquiring 3D annotations is expensive, typically requiring calibrated multi-view setups or labour intensive manual annotations. While annotations of 2D keypoints are much easier to obtain, how to efficiently leverage such *weakly-supervised* data to improve the task of 3D hand pose prediction remains an important open question. The key difficulty stems from the fact that direct application of additional 2D supervision mostly benefits the 2D proxy objective but does little to alleviate the depth and scale ambiguities. Embracing this challenge we propose a set of novel losses that constrain the prediction of a neural network to lie within the range of biomechanically feasible 3D hand configurations. We show by extensive experiments that our proposed constraints significantly reduce the depth ambiguity and allow the network to more effectively leverage additional 2D annotated images. For example, on the challenging freiHAND dataset, using additional 2D annotation without our proposed biomechanical constraints reduces the depth error by only 15%, whereas the error is reduced significantly by 50% when the proposed biomechanical constraints are used.

Keywords: 3D hand pose, weakly-supervised, biomechanical constraints

1 Introduction

Vision-based reconstruction of the 3D pose of human hands is a difficult problem that has applications in many domains. Given that RGB sensors are ubiquitous, recent work has focused on estimating the full 3D pose [6, 19, 26, 36, 46] and dense surface [5, 14, 16] of human hands from 2D imagery alone. This task is challenging due to the dexterity of the human hand, self-occlusions, varying

*This work was done during an internship at NVIDIA.

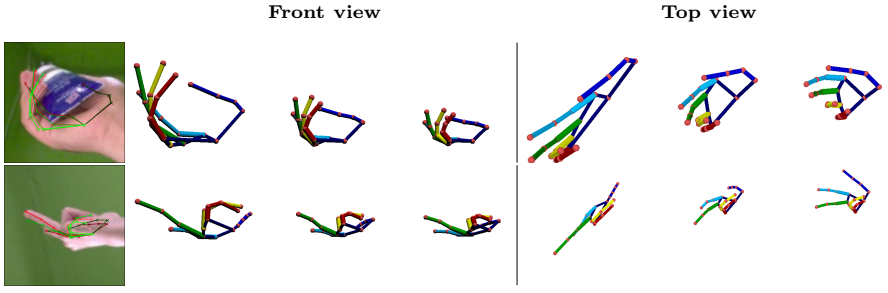


Fig. 1: Impact of the proposed biomechanical constraints (BMC). (b,e) Supplementing fully supervised data with 2D annotated data yields 3D poses with correct 2D projections, yet they are anatomically implausible. (c,f) Adding our biomechanical constraints significantly improves the pose prediction quantitatively and qualitatively. The resulting 3D poses are anatomically valid and display more accurate depth/scale even under severe self- and object occlusions, thus are closer to the ground-truth (d,g).

lighting conditions and interactions with objects. Moreover, any given 2D point in the image plane can correspond to multiple 3D points in world space, all of which project onto that same 2D point. This makes 3D hand pose estimation from monocular imagery an ill-posed inverse problem in which depth and the resulting scale ambiguity pose a significant difficulty.

Most of the recent methods use deep neural networks for hand pose estimation and rely on a combination of fully labeled real and synthetic training data (e.g., [4, 6, 16, 16, 19, 26, 36, 48, 50]). However, acquiring full 3D annotations for real images is very difficult as it requires complex multi-view setups and labour intensive manual annotations of 2D keypoints in all views [15,47,51]. On the other hand, synthetic data does not generalize well to realistic scenarios due to domain discrepancies. Some works attempt to alleviate this by leveraging additional 2D annotated images [5, 19]. Such kind of *weakly-supervised* data is far easier to acquire for real images as compared to full 3D annotations. These methods use these annotations in a straightforward way in the form of a reprojection loss [5] or supervision for the 2D component only [19]. However, we find that the improvements stemming from including the weakly-supervised data in such a manner are mainly a result of 3D poses that agree with the 2D projection. Yet, the uncertainties arising due to depth ambiguities remain largely unaddressed and the resulting 3D poses can still be implausible. Therefore, these methods still rely on large amounts of fully annotated training data to reduce these ambiguities. In contrast, our goal is to *minimize* the requirement of 3D annotated data as much as possible and *maximize* the utility of weakly-labeled real data.

To this end, we propose a set of biomechanically inspired constraints (**BMC**) which can be integrated in the training of neural networks to enable anatomically plausible 3D hand poses even for data with 2D supervision only. Our key insight is that the human hand is subject to a set of limitations imposed by its biomechanics.

We model these limitations in a differentiable manner as a set of *soft constraints*. Note that this is a challenging problem. While the bone length constraints have been used successfully [39, 49], capturing other biomechanical aspects is more difficult. Instead of fitting a hand model to the predictions, we *extract* the quantities in question directly from the predictions to impose our constraints. As such, the method of extraction has to be carefully designed to work under noisy and malformed 3D joint predictions while simultaneously being fully differentiable under any pose. We propose to encode these constraints into a set of losses that are fully differentiable, interpretable and which can be incorporated into the training of any deep learning architecture that predicts 3D joint configurations. Due to this integration, we do not require a post-refinement step during test time. More specifically, our set of soft constraints consists of three equations that define i) the range of valid bone lengths, ii) the range of valid palm structure, and iii) the range of valid joint angles of the thumb and fingers. The main advantage of our set of constraints is that all parameters are interpretable and can either be set manually, opening up the possibility of personalization, or be obtained from a small set of data points for which 3D labels are available. As backbone model, we use the 2.5D representation proposed by Iqbal *et al.* [19] due to its superior performance. We identify an issue in absolute depth calculation and remedy it via a novel refinement network. In summary, we contribute:

- A novel set of differentiable soft constraints inspired by the biomechanical structure of the human hand.
- Quantitative and qualitative evidence that demonstrates that our proposed set of constraints improves 3D prediction accuracy in *weakly supervised settings*, resulting in an improvement of 55% as opposed to 32% as yielded by straightforward use of weakly-supervised data.
- A neural network architecture that extends [19] with a refinement step.
- Achieving state-of-the-art performance on Dexter+Object using only synthetic and weakly-supervised real data, indicating cross-data generalizability.

The proposed constraints require no special data nor are they specific to a particular backbone architecture.

2 Related work

Hand pose estimation from monocular RGB has gained traction in recent years due numerous possible applications. Generally there are two trains of thought.

Model-based methods ensure plausible poses by fitting a hand model to the observation via optimization. As they are not learning-based, they are sensitive to initial conditions, rely on temporal information [18, 27–29] or do not take the image into consideration during optimization [29]. Whereas some make use of geometric primitives [27–29], other simply model the joint angles directly [8,11,21,23,32,43], learn a lower dimensional embedding of the joints [24], pose [18] or go a step further and model muscles of the hand [1]. Different to these methods, we propose to incorporate these constraints *directly* into the training procedure of a neural network in a fully differentiable manner. As such,

we do not fit a hand model to the prediction, but *extract* and constrain the biomechanical quantities from them directly. The resulting network predicts biomechanically-plausible poses and does not suffer from the same disadvantages.

Learning-based methods utilize neural networks that either directly regress the 3D positions of the hand keypoints [19, 26, 36, 40, 46, 50] or predict the parameters of a deformable hand model [4, 5, 16, 44, 48]. Zimmermann *et al.* [50] are the first to use deep neural network for root-relative 3D hand pose estimation from RGB images via a multi-staged approach. Spurr *et al.* [36] learn a unified latent space that projects multiple modalities into the same space, learning a lower level embedding of the hands. Similarly, Yang *et al.* [46] learn a latent space that disentangles background, camera and hand pose. However, all these methods require large numbers of fully labeled training data. Cai *et al.* [6] try to alleviate this problem by introducing an approach that utilizes paired RGB-D images to regularize the depth predictions. Mueller *et al.* [26] attempt to improve the quality of synthetic training data by learning a GAN model that minimizes the discrepancies between real and synthetic images. Iqbal *et al.* [19] decompose the task into learning 2D and root-relative depth components. This decomposition allows to use weakly-labeled real images with only 2D pose annotations which are cheap to acquire. While these methods demonstrate better generalization by adding a large number weakly-labeled training samples, the main drawback of this approach is that the depth ambiguities remain unaddressed. As such, training using only 2D pose annotations does not impact the depth predictions. This may result in 3D poses with accurate 2D projections, but due to depth ambiguities the 3D poses can still be implausible. In contrast, in this work, we propose a set of biomechanical constraints that ensures that the predicted 3D poses are always anatomically plausible during training (see Fig. 1). We formulate these constraints in form of a fully-differentiable loss functions which can be incorporated into any deep learning architecture that predicts 3D joint configurations. We use a variant of Iqbal *et al.* [19] as a baseline and demonstrate that the requirement of fully labeled real images can be significantly minimized while still maintaining performance on par with fully-supervised methods.

Other recent methods directly predict the parameters of a deformable hand model, *e.g.*, MANO [33], from RGB images [5, 16, 30, 44, 48]. The predicted parameters consist of the shape and pose deformations wrt. a mean shape and pose that are learned using large amounts of 3D scans of the hand. Alternatively, [14, 22] circumvent the need for a parametric hand model by directly predicting the mesh vertices from RGB images. These methods require both shape and pose annotations for training, therefore obtaining such kind of training data is even harder. Hence, most methods rely on synthetic training data. Some methods [4, 5, 48] alleviate this by introducing re-projection losses that measure the discrepancy between the projection of 3D mesh with labeled 2D poses [5] or silhouettes [4, 48]. Even though they utilize strong hand priors in form of a mean hand shape and by operating on a low-dimensional PCA space, using re-projection losses with weakly-labeled data still does not guarantee that the resulting 3D poses will be anatomically plausible. Therefore, all these methods

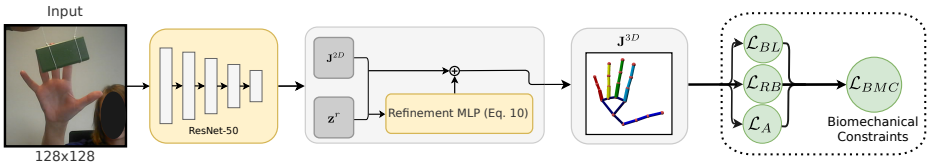


Fig. 2: Method overview. A model takes an RGB image and predicts the 3D joints on which we apply our proposed BMC. These guide the model to predict plausible poses.

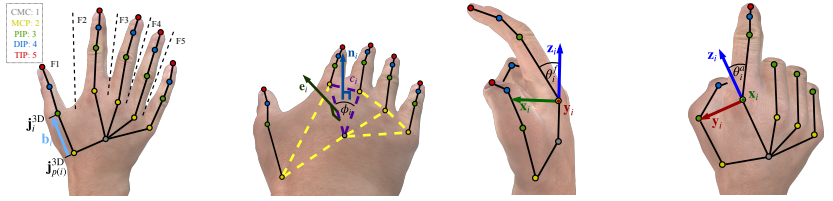
rely on a large number of fully labeled training data. In body pose estimation, such methods generally resort to adversarial losses to ensure plausibility [20].

Biomechanical constraints have also been used in the literature to encourage plausible 3D poses by imposing biomechanical limits on the structure of the hands [9, 10, 12, 25, 34, 38, 41, 42, 45] or via a learned refinement model [7]. Most methods [2, 9, 10, 25, 34, 38, 41, 45] impose these limits via inverse kinematic in a post-processing step, therefore the possibility of integrating them for neural network training remains unanswered. Our proposed soft-constraints are fully integrated into the network, which does not require a post-refinement step during test time. Similar to our method, [12, 42] also penalize invalid bone lengths. However, we additionally model the joint limits and palmar structure.

3 Method

Our method is summarized in Figure 2. Our key contribution is a set of novel constraints that constitute a biomechanical model of the human hand and capture the bone lengths, joint angles and shape of the palm. We emphasize that we do not fit a kinematic model to the predictions, but instead extract the quantities in question directly from the predictions in order to constrain them. Therefore the method of extraction is carefully designed to work under noisy and malformed 3D joint predictions while simultaneously being fully differentiable in any configuration. These biomechanical constraints provide an inductive bias to the neural network. Specifically, the network is guided to predict anatomically plausible hand poses for weakly-supervised data (*i.e.* 2D only), which in turn increases generalizability. The model can be combined with any backbone architecture that predicts 3D keypoints. We first introduce the notations used in this paper followed by the details of the proposed biomechanical losses. Finally, we discuss the integration with a variant of [19].

Notation. We use bold capital font for matrices, bold lowercase for vector and roman font for scalars. We assume a right hand. The joints $[\mathbf{j}_1^{3D}, \ldots, \mathbf{j}_{21}^{3D}] = \mathbf{J}^{3D} \in \mathbb{R}^{21\times 3}$ define a kinematic chain of the hand starting from the root joint $\mathbf{j}_1^{3D}$ and ending in the fingertips. For the sake of simplicity, the joints of the hands are grouped by the fingers, denoted as the respective set $F1, \ldots, F5$, visualized in Fig. 3a. Each $\mathbf{j}_i^{3D}$, except the root joint (CMC), has a parent, denoted as $p(i)$. We define a bone $\mathbf{b}_i = \mathbf{j}_{i+1}^{3D} - \mathbf{j}_{p(i+1)}^{3D}$ as the vector pointing



a) Joint skeleton structure b) Root bone structure c) Angles. Flexion: Left – Abduction: Right

Fig. 3: Illustration of our proposed biomechanical structure.

from the parent joint to its child joint. Hence $[\mathbf{b}_1, \ldots, \mathbf{b}_{20}] = \mathbf{B} \in \mathbb{R}^{20\times 3}$. The bones are named according to the child joint. For example, the bone connecting MCP to PIP is called PIP bone. We define the five root bones as the MCP bones, where one endpoint is the root $\mathbf{j}_1^{3D}$. Intuitively, the root bones are those that lie within and define the palm. We define the bones $\mathbf{b}_i$ with $i = 1, \ldots, 5$ to correspond to the root bones of fingers $F1, \ldots, F5$. We denote the angle $\alpha(v_1, v_2) = \arccos\left(\frac{\mathbf{v}_1^T \mathbf{v}_2}{||\mathbf{v}_1||_2 \, ||\mathbf{v}_2||_2}\right)$ between the vectors $\mathbf{v}_1, \mathbf{v}_2$. The interval loss is defined as $\mathcal{I}(x; a, b) = \max(a - x, 0) + \max(x - b, 0)$. The normalized vector is defined as $\text{norm}(\mathbf{x}) = \frac{\mathbf{x}}{||\mathbf{x}||_2}$. Lastly, $\mathrm{P}_{\mathbf{xy}}(\mathbf{v})$ is the orthogonal projection operator, projecting $\mathbf{v}$ orthogonally onto the $\mathbf{x}$-$\mathbf{y}$ plane where $\mathbf{x}$,$\mathbf{y}$ are vectors.

3.1 Biomechanical constraints

Our goal is to integrate our biomechanical soft constraints (BMC) into the training procedure that encourages the network to predict feasible hand poses. We seek to avoid iterative optimization approaches such as inverse kinematics in order to avert significant increases in training time.

The proposed model consists of three functional parts, visualized in Fig. 3. First, we consider the length of the bones, including the root bones of the palm. Second, we model the structure and shape of the palmar region, consisting of a rigid structure made up of individual joints. To account for inter-subject variability of bones and palm structure, it is important to not enforce a specific mean shape. Instead, we allow for these properties to lie within a valid range. Lastly, the model describes the articulation of the individual fingers. The finger motion is described via modeling of the flexion and abduction of individual bones. As their limits are interdependent, they need to be modeled jointly. As such, we propose a novel constraint that takes this interdependence into account.

The limits for each constraint can be attained manually from measurements, from the literature (e.g [9, 34]), or acquired in a data-driven way from 3D annotations, should they be available.

Bone length. For each bone i, we define an interval $[b_i^{\min}, b_i^{\max}]$ of valid bone length and penalize if the length $||\mathbf{b}_i||_2$ lies outside of this interval:

$$\mathcal{L}_{\text{BL}}(\mathbf{J}^{3D}) = \frac{1}{20} \sum_{i=1}^{20} \mathcal{I}(||\mathbf{b}_i||_2; b_i^{\min}, b_i^{\max})$$

This loss encourages keypoint predictions that yield valid bone lengths. Fig. 3a shows the length of a bone in blue.

Root bones. To attain valid palmar structures we first interpret the root bones as spanning a mesh and compute its curvature by following [31]:

$$c_i = \frac{(\mathbf{e}_{i+1} - \mathbf{e}_i)^T(\mathbf{b}_{i+1} - \mathbf{b}_i)}{||\mathbf{b}_{i+1} - \mathbf{b}_i||^2}, \text{ for } i \in \{1,2,3,4\} \tag{1}$$

Where $\mathbf{e}_i$ is the edge normal at bone $\mathbf{b}_i$:

$$\begin{aligned} \mathbf{n}_i &= \text{norm}(\mathbf{b}_{i+1} \times \mathbf{b}_i), \text{for } i \in \{1,2,3,4\} \\ \mathbf{e}_i &= \begin{cases} \mathbf{n}_1, & \text{if } i = 1 \\ \text{norm}(\mathbf{n}_i + \mathbf{n}_{i-1}), & \text{if } i \in \{2,3,4\} \\ \mathbf{n}_4, & \text{if } i = 5 \end{cases} \end{aligned} \tag{2}$$

Positive values of c_i denote an arched hand, for example when pinky and thumb touch. A flat hand has no curvature. Fig. 3b visualizes the mesh in dashed yellow and the triangle over which the curvature is computed in dashed purple.

We ensure that the root bones fall within correct angular ranges by defining the angular distance between neighbouring $\mathbf{b}_i$,$\mathbf{b}_{i+1}$ across the plane they span:

$$\phi_i = \alpha(\mathbf{b}_i, \mathbf{b}_{i+1}) \tag{3}$$

We constrain both the curvature c_i and angular distance ϕ_i to lie within a valid range $[c_i^{\min}, c_i^{\max}]$ and $[\phi_i^{\min}, \phi_i^{\max}]$:

$$\mathcal{L}_{\text{RB}}(\mathbf{J}^{3D}) = \frac{1}{4}\sum_{i=1}^{4}\left(\mathcal{I}(c_i; c_i^{\min}, c_i^{\max}) + \mathcal{I}(\phi_i; \phi_i^{\min}, \phi_i^{\max})\right)$$

$\mathcal{L}_{\text{RB}}$ ensures that the predicted joints of the palm define a valid structure, which is crucial since the kinematic chains of the fingers originate from this region.

Joint angles. To compute the joint angles, we first need to define a consistent frame $\mathbf{F}_i$ of a local coordinate system for each finger bone $\mathbf{b}_i$. $\mathbf{F}_i$ must be consistent with respect to the movements of the finger. In other words, if one constructs $\mathbf{F}_i$ given a pose $\mathbf{J}_1^{3D}$, then moves the fingers and corresponding $\mathbf{F}_i$ into pose $\mathbf{J}_2^{3D}$, the resulting $\mathbf{F}_i$ should be the same as if constructed from $\mathbf{J}_2^{3D}$ directly.

We assume right-handed coordinate systems. To construct $\mathbf{F}_i$, we define two out of three axes based on the palm. We start with the first layer of fingers bones (PIP bones). We define their respective z-component of $\mathbf{F}_i$ as the normalized bone of their respective parent bone (in this case, the root bones): $\mathbf{z}_i = \text{norm}(\mathbf{b}_{p(i)})$. Next, we define the x-axis, based on the plane normals spanned by two neighbouring root bones:

$$\mathbf{x}_i = \begin{cases} -\mathbf{n}_{p(i)}, & \text{if } p(i) \in \{1,2\} \\ -\text{norm}(\mathbf{n}_{p(i)} + \mathbf{n}_{p(i)-1}), & \text{if } p(i) \in \{3,4\} \\ -\mathbf{n}_4, & \text{if } p(i) = 5 \end{cases} \tag{4}$$

Where $\mathbf{n}_i$ is defined as in Eq. 2. Lastly, we compute the last axis $\mathbf{y}_i = \text{norm}(\mathbf{z}_i \times \mathbf{x}_i)$. Given $\mathbf{F}_i$, we can now define the flexion and abduction angles. Each of these angles are given with respect to the local z-axis of $\mathbf{F}_i$. Given $\mathbf{b}_i$ in its local coordinates $\mathbf{b}_i^{\mathbf{F}_i}$ wrt. $\mathbf{F}_i$, we define the flexion and abduction angles as:

$$\begin{aligned} \theta_i^{\text{f}} &= \alpha(\text{P}_{xz}(\mathbf{b}_i^{\mathbf{F}_i}), \mathbf{z}_i) \\ \theta_i^{\text{a}} &= \alpha(\text{P}_{xz}(\mathbf{b}_i^{\mathbf{F}_i}), \mathbf{b}_i^{\mathbf{F}_i}) \end{aligned} \tag{5}$$

Fig. 3c visualizes $\mathbf{F}_i$ and the resulting angles. Note that this formulation leads to ambiguities, where different bone orientations can map to the same $(\theta_i^{\text{f}}, \theta_i^{\text{a}})$-point. We resolve this via an octant lookup, which leads to angles in the intervals $\theta_i^{\text{f}} \in [-\pi, \pi]$ and $\theta_i^{\text{a}} \in [-\pi/2, \pi/2]$ respectively. See appendix for more details.

Given the angles of the first set of finger bones, we can then construct the remaining two rows of finger bones. Let $\mathbf{R}^{\theta_i}$ denote the rotation matrix that rotates by θ_i^{f} and θ_i^{a} such that $\mathbf{R}^{\theta_i}\mathbf{z}_i = \mathbf{b}_i^{\mathbf{F}_i}$, then we iteratively construct the remaining frames along the kinematic chain of the fingers:

$$\mathbf{F}_i = \mathbf{R}^{\theta_i}\mathbf{F}_{p(i)} \tag{6}$$

This method of frame construction via rotating by θ_i^{f} and θ_i^{a} ensures consistency across poses. The remaining angles can be acquired as described in Eq. 5.

Lastly, the angles need to be constrained. One way to do this is to consider each angle independently and penalize them if they lie outside an interval. This corresponds to constraining them within a box in a 2D space, where the endpoints are the min/max of the limits. However, finger angles have inter-dependency, therefore we propose an alternative approach to account for this. Given points $\theta_i = (\theta_i^{\text{f}}, \theta_i^{\text{a}})$ that define a range of motion, we approximate their convex hull on the $(\theta^{\text{f}}, \theta^{\text{a}})$-plane with a fixed set of points $\mathcal{H}_i$. The angles are constrained to lie within this structure by minimizing their distance to it:

$$\mathcal{L}_{\text{A}}(\mathbf{J}^{3D}) = \frac{1}{15}\sum_{i=1}^{15} D_H(\theta_i, \mathcal{H}_i) \tag{7}$$

Where D_H is the distance of point θ_i to the hull $\mathcal{H}_i$. Details on the convex hull approximation and implementation can be found in the appendix.

3.2 $\mathbf{Z}^{\mathbf{root}}$ Refinement

The 2.5D joint representation allows us to recover the value of the absolute pose Z^{root} up to a scaling factor . This is done by solving a quadratic equation dependent on the 2D projection $\mathbf{J}^{2D}$ and relative depth values $\mathbf{z}^r$, as proposed in [19]. In practice, small errors in $\mathbf{J}^{2D}$ or $\mathbf{z}^r$ can result in large deviations of Z^{root}. This leads to big fluctuations in the translation and scale of the predicted pose, which is undesirable. To alleviate these issues, we employ an MLP to refine and smooth the calculated $\hat{Z}^{root}$:

$$\hat{Z}^{root}_{\text{ref}} = \hat{Z}^{root} + M_{\text{MLP}}(\mathbf{z}^r, \mathbf{K}^{-1}\mathbf{J}^{2D}, \hat{Z}^{root}; \omega) \tag{8}$$

Where M_{MLP} is a multilayered perceptron with parameters ω that takes the predicted and calculated values $\mathbf{z}^r \in \mathbb{R}^{21}$, $\mathbf{K}^{-1}\mathbf{J}^{2D} \in \mathbb{R}^{21\times 3}$, $Z^{root} \in \mathbb{R}$ and outputs a residual term. Alternatively, one could predict Z^{root} directly using an MLP with the same input. However, as the exact relationship between the predicted variables and Z^{root} is known, we resort to the refinement approach instead of requiring a model to learn what is already known.

3.3 Final loss

The biomechanical soft constraints is constructed as follows:

$$\mathcal{L}_{\text{BMC}} = \lambda_{\text{BL}}\mathcal{L}_{\text{BL}} + \lambda_{\text{RB}}\mathcal{L}_{\text{RB}} + \lambda_{\text{A}}\mathcal{L}_{\text{A}} \tag{9}$$

Our final model is trained on the following loss function:

$$\mathcal{L} = \lambda_{\mathbf{J}^{2D}}\mathcal{L}_{\mathbf{J}^{2D}} + \lambda_{\mathbf{z}^r}\mathcal{L}_{\mathbf{z}^r} + \lambda_{\text{Z}^{\text{root}}_{\text{ref}}}\mathcal{L}_{\text{Z}^{\text{root}}} + \mathcal{L}_{\text{BMC}} \tag{10}$$

Where $\mathcal{L}_{\mathbf{J}^{2D}}$, $\mathcal{L}_{\mathbf{z}^r}$ and $\mathcal{L}_{\text{Z}^{\text{root}}}$ are the L1 loss on any available $\mathbf{J}^{2D}$, $\mathbf{z}^r$ and Z^{root} labels respectively. The weights λ_i balance the individual loss terms.

4 Implementation

We use a ResNet-50 backbone [17]. The input to our model is a 128×128 RGB image from which the 2.5D representation is directly regressed. The model and its refinement step is trained on fully supervised and weakly-supervised data. The network was trained for 70 epochs using SGD with a learning rate of $5e$–3 and a step-wise learning rate decay of 0.1 after every 30 epochs. We apply the biomechanical constraints directly on the predicted 3D keypoints $\mathbf{J}^{3D}$.

5 Evaluation

Here we introduce the datasets used, show the performance of our proposed $\mathcal{L}_{\mathbf{BMC}}$ and compare in extensive settings. Specifically, we study the effect of adding weakly supervised data to complement fully supervised training. All experiments are conducted in a setting where we assume access to a fully supervised dataset, as well as a supplementary weakly supervised real dataset. Therefore we have access to 2D ground-truth annotations and the computed constraint limits. We study two cases of 3D supervision sources:

Synthetic data. We choose RHD. Acquiring fully labeled synthetic data is substantially easier as compared to real data. Section 5.3-5.5 consider this setting.

Partially labeled real data. In Section 5.6 we gradually increase the number of real 3D labeled samples to study how the proposed approach works under different ratio of fully to weakly supervised data.

To make clear what kind of supervision is used we denote $\mathbf{3D}_{\text{A}}$ if 3D annotation is used from dataset A. We indicate usage of 2D from dataset A as $\mathbf{2D}_{A}$. Section 5.3 and 5.4 are evaluated on FH.

5.1 Datasets

Each dataset that provides 3D labels comes with the camera intrinsics. Hence the 2D pose can be easily acquired from the 3D pose. Tab. 1 provides an overview of datasets used. The test set of HO-3D and FH are available only via a submission system with limited number of total submissions. Therefore for the ablation study (Section 5.4) and inspecting the effect of weak-supervision (Section 5.3), we divide the training set into a training and validation split. For these sections, we choose to evaluate on FH due to its large number of samples and variability in both hand pose and shape.

Table 1: Overview of datasets used for evaluation.

Name	Type	joints #	train/test #
Rendered Hand Pose (RHD) [50]	Synth	21	42k / 2.7k
FreiHAND (FH) [51]	Real	21	33k / 4.0k
Dexter+Object (D+O) [37]	Real	5	- / 3.1k
Hand-Object 3D (HO-3D) [15]	Real	21	11k / 6.6k

5.2 Evaluation Metric

HO-3D. The error given by the submission system is the mean joint error in mm. The INTERP is the error on test frames sampled from training sequences that are not present in the training set. The EXTRAP is the error on test samples that have neither hand shapes nor objects present in the training set. We used the version of the dataset that was available at the time [3].
FH. The error given by the submission system is the mean joint error in mm. Additionally, the area under the curve (AUC) of the percentage of correct keypoints (PCK) plot is reported. The PCK values lie in an interval from 0 mm to 50 mm with 100 equally spaced thresholds. Both the aligned (using procrustes analysis) and unaligned scores are given. We report the aligned score. The unaligned score can be found in the appendix.
D+O. We report the AUC for the PCK thresholds of 20 to 50 mm comparable with prior work [5, 48, 51]. For [19, 26, 36, 50] we report the numbers as presented in [48] as they consolidate all AUC of related work in a consistent manner using the same PCK thresholds. For [4], we recomputed the AUC for the same interval based on the values provided by the authors.

5.3 Effect of Weak-Supervision

We first inspect how weak-supervision affects the performance of the model. We decompose the 3D prediction error on the validation set of FH in terms of its 2D ($\mathbf{J}^{2D}$) and depth component ($\mathbf{Z}$) via the pinhole camera model $\mathbf{Z}^{-1}\mathbf{K}\mathbf{J}^{3D} = \mathbf{J}^{2D}$ and evaluate their individual error.

We train four models using different data sources. 1) Full 3D supervision on both synthetic RHD and real FH ($\mathbf{3D}_{\text{RHD}} + \mathbf{3D}_{\text{FH}}$), which serves as an upper bound for when all 3D labels are available 2) Fully supervised on RHD which constitutes our lower bound on accuracy ($\mathbf{3D}_{\text{RHD}}$) 3) Fully supervised on RHD with naive application of weakly-supervised FH ($+\mathbf{2D}_{\text{FH}}$) 4) Like setting 3) but adding our proposed constraints ($+\mathcal{L}_{\mathbf{BMC}}$).

Table 2: The effect of weak-supervision on the **validation** split of FH. Training on synthetic data (RHD) leads to poor accuracy on real data (FH). Adding real 2D labeled data reduces 3D prediction error due to better alignment with the 2D projection. Adding our proposed $\mathcal{L}_{\mathbf{BMC}}$ significantly reduces the 3D error due to more accurate **Z**.

Effect of weak-supervision	Description	Mean Error ↓ 2D (px)	Z (mm)	3D (mm)
$\mathbf{3D}_{\mathrm{RHD}} + \mathbf{3D}_{\mathrm{FH}}$	Fully supervised, synthetic+real	3.72	5.69	8.78
$+ \mathcal{L}_{\mathbf{BMC}}$ (**ours**)	+ BMC	**3.70**	**5.44**	**8.60**
$\mathbf{3D}_{\mathrm{RHD}}$	Fully supervised, synthetic only	12.35	20.02	30.82
$+ \mathbf{2D}_{\mathrm{FH}}$	+ Weakly supervised, real	3.80	17.02	20.92
$+ \mathcal{L}_{\mathbf{BMC}}$ (**ours**)	+ BMC	**3.79**	**9.97**	**13.78**

Tab. 2 shows the results. The model trained with full 3D supervision from real and synthetic data reflects the best setting. Adding $\mathcal{L}_{\mathbf{BMC}}$ during training slightly reduces 3D error (8.78mm to 8.6mm) primarily due to a regularization effect. When the model is trained only on synthetic data ($\mathbf{3D}_{\mathrm{RHD}}$) we observe a significant rise (8.78mm to 30.82mm) in 3D error due to the poor generalization from synthetic data. When weak-supervision is provided from the real data ($+\mathbf{2D}_{\mathrm{FH}}$), the error is reduced (30.82mm to 20.92mm). However, inspecting this more closely we observe that the improvement comes mainly from 2D error reduction (12.35px to 3.8px), whereas the depth component is improved marginally (20.02mm to 17.02mm). Observing these samples qualitatively (Fig. 1), we see that many do not adhere to biomechanical limits of the human hand. By penalizing such violations via our proposed losses $\mathcal{L}_{\mathbf{BMC}}$ to the weakly supervised setting we see a significant improvement in 3D error (20.92mm to 13.78mm) which is due to improved depth accuracy (20.02mm to 9.97mm). Inspecting (*e.g.* Fig. 1) closer, we see that the model predicts the correct 3D pose in challenging settings such as heavy self- and object occlusion, despite having never seen such samples in 3D. Since $\mathcal{L}_{\mathbf{BMC}}$ describes a valid range, rather than a specific pose, slight deviations from the ground truth 3D pose have to be expected which explains the small remaining quantitative gap from the fully supervised model.

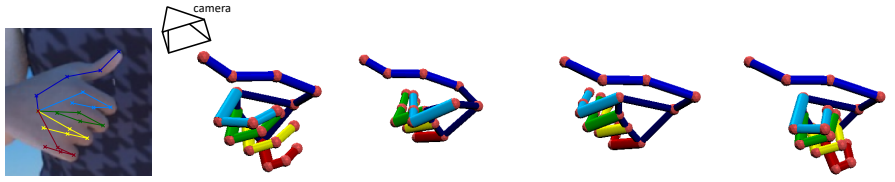
5.4 Ablation Study

We quantify the individual contributions of our proposals on the validation set of FH and reproduce these results on HO-3D in supplementary. Each error metric is computed for the root-relative 3D pose.

Refinement network. Tab. 3 shows the impact of Z^{root} refinement (Sec. 3.2). We train two models that include (w. refinement) or omit (w/o refinement) the refinement step, using full supervision on FH ($\mathbf{3D}_{\mathrm{FH}}$). Using refinement, the mean error is reduced by 1.44mm which indicates that refining effectively reduces outliers.

Table 3: Effect of Z^{root} refinement

Ablation Study	EPE (mm) mean ↓	median ↓	AUC ↑
w/o refinement	11.20	8.62	0.95
w. refinement (**ours**)	**9.76**	**8.14**	**0.97**



(a) Input image (b) Ground-truth (c) $\mathcal{L}_{\text{BL}}$ (d) $\mathcal{L}_{\text{BL}} + \mathcal{L}_{\text{RB}}$ (e) $\mathcal{L}_{\text{BL}} + \mathcal{L}_{\text{RB}} + \mathcal{L}_{\text{A}}$

Fig. 4: Impact of our proposed losses. (a) All predicted 3D poses project to the same 2D pose. (b) Ground-truth pose. (c) $\mathcal{L}_{\text{BL}}$ results in poses that have correct bone lengths, but may have invalid angles and palm structure. (d) Including $\mathcal{L}_{\text{RB}}$ imposes a correct palm, but the fingers are still articulated wrong. (e) Adding $\mathcal{L}_{\text{A}}$ leads to the finger bones having correct angles. The resulting hand is plausible and close to the ground-truth.

Components of BMC. In Tab. 4, we perform a series of experiments where we incrementally add each of the proposed constraints. For 3D guidance, we use the synthetic RHD and *only* use the 2D labels of FH. We first run the baseline model trained only on this data ($\mathbf{3D}_{\text{RHD}} + \mathbf{2D}_{\text{FH}}$). Next, we add the bone length loss $\mathcal{L}_{\text{BL}}$, followed by the root bone loss $\mathcal{L}_{\text{RB}}$ and the angle loss $\mathcal{L}_{\text{A}}$. An upper bound is given by our model trained fully supervised on both datasets ($\mathbf{3D}_{\text{RHD}} + \mathbf{3D}_{\text{FH}}$). Each component contributes positively towards the final performance, totalling a decrease of 6.24mm in mean error as compared to our weakly-supervised baseline, significantly closing the gap to the fully supervised upper bound. A qualitative assessment of the individual losses can be seen in Fig. 4.

Table 4: Effect of BMC components.

Ablation Study	EPE (mm) mean ↓	EPE (mm) median ↓	AUC ↑
$\mathbf{3D}_{\text{RHD}} + \mathbf{2D}_{\text{FH}}$	20.92	16.93	0.81
+ $\mathcal{L}_{\text{BL}}$ (**ours**)	17.58	14.81	0.88
+ $\mathcal{L}_{\text{RB}}$ (**ours**)	15.48	13.49	0.91
+ $\mathcal{L}_{\text{A}}$ (**ours**)	**13.78**	**11.61**	**0.92**
$\mathbf{3D}_{\text{RHD}} + \mathbf{3D}_{\text{FH}}$	8.78	7.25	0.98

Co-dependency of angles. In Tab. 5, we show the importance of modeling the dependencies between the flexion and abduction angle limits (Sec. 3), instead of regarding them independently. Co-dependent angle limits yield a decrease in mean error of 1.40 mm.

Table 5: Effect of angle constraints

Ablation Study	EPE (mm) mean ↓	EPE (mm) median ↓	AUC ↑
Independent	15.57	13.45	0.91
Dependent	13.78	11.61	0.92

Constraint limits. In Tab. 6, we investigate the effect of the used limits on the final performance, as one may have to resort to approximations. For this, we instead take the hand parameters from RHD and perform the same weakly-supervised experiment as before ($+\mathcal{L}_{\mathbf{BMC}}$). Approximating the limits from another dataset slightly increases the error, but still clearly outperforms the 2D baseline.

Table 6: Effect of limits

Ablation Study	EPE (mm) mean ↓	EPE (mm) median ↓	AUC ↑
Approximated	16.14	13.93	0.90
Computed	13.78	11.61	0.92

Table 7: Results on the respective **test** split, evaluated by the *submission systems*. Training on RHD leads to poor accuracy on both FH and HO-3D. Adding weakly-supervised data improves results, as expected. By including our proposed $\mathcal{L}_{\mathbf{BMC}}$, our model incurs a significant boost in accuracy, especially evident for the INTERP score.

	Description	R=FH		R=HO-3D	
		mean ↓	AUC ↑	EXTRAP ↓	INTERP ↓
$\mathbf{3D}_{\text{RHD}} + \mathbf{3D}_{\text{R}}$	Fully sup. upper bound	0.90	0.82	18.22	5.02
$\mathbf{3D}_{\text{RHD}}$	Fully sup. lower bound	1.60	0.69	20.84	33.57
$+\mathbf{2D}_{\text{R}}$	+ Weakly sup.	1.26	0.75	19.57	25.16
$+ \mathcal{L}_{\mathbf{BMC}}$ (**ours**)	+ BMC	**1.13**	**0.78**	**18.42**	**10.31**

Table 8: Datasets used by prior work for evaluation on D+O. With solely fully-supervised synthetic and weakly-supervised real data, we outperform recent works and perform on par with [48]. All other works rely on full supervision from real and synthetic data. *These works report unaligned results.

D+O	Annotations used			AUC ↑
	Synth.	Real	Scans	
Ours (weakly sup.)	3D	**2D only**		**0.82**
Zhang (2019) [48]	3D	3D	3D	0.82
Boukhayma (2019) [5]	3D	3D	3D	0.76
Iqbal (2018)* [19]	3D	3D		0.67
Baek (2019)* [4]	3D	3D	3D	0.61
Zimmermann (2018) [50]	3D	3D		0.57
Spurr (2018) [36]	3D	3D		0.51
Mueller (2018)* [26]	3D	Unlabeled		0.48

5.5 Bootstrapping with Synthetic Data

We validate $\mathcal{L}_{\mathbf{BMC}}$ on the **test set** of FH and HO-3D. We train the same four models like in Sec. 5.3 using fully supervised RHD and weakly-supervised real data R∈[FH,HO-3D].

For all results here we perform training on the *full* dataset and evaluate on the official test split via the online submission system. Additionally, we evaluate the cross-dataset performance on D+O dataset to show how our proposed constraints improves generalizability and compare with prior work [4, 5, 19, 26, 48].

FH. The second column of Tab. 7 shows the dataset performance for R = FH. Training solely on RHD ($\mathbf{3D}_{\text{RHD}}$) performs the worst. Adding real data ($+\mathbf{2D}_{\text{FH}}$) with 2D labels reduces the error, as we reduce the real/synthetic domain gap. Including the proposed $\mathcal{L}_{\mathbf{BMC}}$ results in an accuracy boost.

HO-3D. The third column of Tab. 7 shows a similar trend for R = HO-3D. Most notably, our constraints yield a decrease of 14.85 mm for INTERP. This is significantly larger than the relative decrease the 2D data adds (-8.41mm). For EXTRAP, BMC yields an improvement of 1.15mm, which is close to the 1.27mm gained from 2D data. This demonstrates that $\mathcal{L}_{\mathbf{BMC}}$ is beneficial in leveraging 2D data more effectively in unseen scenarios.

D+O. In Tab. 8 we demonstrate the cross-data performance on D+O for R = FH. Most recent works have made use of MANO [4, 5, 48], leveraging a low-dimensional embedding of highly detailed hand scans and require custom

synthetic data [4,5] to fit the shape. Using only fully supervised *synthetic data* and *weakly-supervised* real data in conjunction with $\mathcal{L}_{\mathbf{BMC}}$, we reach state-of-the-art.

5.6 Bootstrapping with Real Data

We study the impact of our biomechanical constrains on reducing the number of labeled samples required in scenarios where few real 3D labeled samples are available. We train a model in a setting where a fraction of the data contains the full 3D labels and the remainder contains only 2D supervision.

Here we choose $R = \text{FH}$, use the entire training set and evaluate on the test set. For each fraction of fully labelled data we evaluate two models. The first is trained on both the fully and weakly labeled samples. The second is trained with the addition of our proposed constraints. We show the results in Fig. 5. For a given AUC, we plot the number of labeled samples required to reach it. We observe that for lower labeling percentages, the amount of labeled data required is approximately *half* using $\mathcal{L}_{\mathbf{BMC}}$. This showcases its effectiveness in low label settings and demonstrates the decrease in requirement for fully annotated training data.

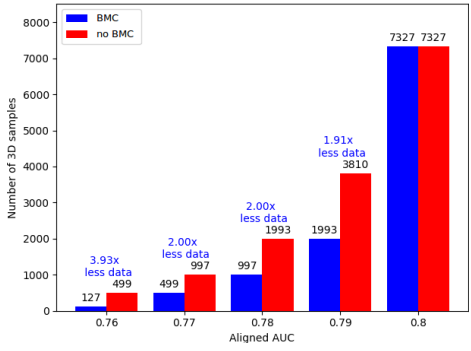


Fig. 5: Number of 3D samples required to reach a certain aligned AUC on FH.

6 Conclusion

We propose a set of fully differentiable biomechanical losses to more effectively leverage weakly supervised data. Our method consists of a novel procedure to encourage anatomically correct predictions of a backbone network via a set of novel losses that penalize invalid bone length, joint angles as well as palmar structures. Furthermore, we have experimentally shown that our constraints can more effectively leverage weakly-supervised data, which show improvement on both within- and cross-dataset performance. Our method reaches state-of-the-art performance on the aligned D+O objective using 3D synthetic and 2D real data and reduces the need of training data by half in low label settings on FH.

Acknowledgments. We are grateful to Christoph Gebhardt and Shoaib Ahmed Siddiqui for the aid in figure creation and Abhishek Badki for helpful discussions.

References

1. Albrecht, I., Haber, J., Seidel, H.P.: Construction and animation of anatomically based human hand models. In: SIGGRAPH (2003)
2. Aristidou, A.: Hand tracking with physiological constraints. The Visual Computer **34**(2), 213–228 (2018)
3. Armagan, A., Garcia-Hernando, G., Baek, S., Hampali, S., Rad, M., Zhang, Z., Xie, S., Chen, M., Zhang, B., Xiong, F., et al.: Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3d hand pose estimation under hand-object interaction. In: ECCV (2020)
4. Baek, S., Kim, K.I., Kim, T.K.: Pushing the envelope for RGB-based dense 3d hand pose estimation via neural rendering. In: CVPR (2019)
5. Boukhayma, A., Bem, R.d., Torr, P.H.: 3d hand shape and pose from images in the wild. In: CVPR (2019)
6. Cai, Y., Ge, L., Cai, J., Yuan, J.: Weakly-supervised 3d hand pose estimation from monocular RGB images. In: ECCV (2018)
7. 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: CVPR (2019)
8. Cerveri, P., De Momi, E., Lopomo, N., Baud-Bovy, G., Barros, R., Ferrigno, G.: Finger kinematic modeling and real-time hand motion estimation. Annals of biomedical engineering **35**(11), 1989–2002 (2007)
9. Chen Chen, F., Appendino, S., Battezzato, A., Favetto, A., Mousavi, M., Pescarmona, F.: Constraint study for a hand exoskeleton: human hand kinematics and dynamics. Journal of Robotics (2013)
10. Cobos, S., Ferre, M., Uran, M.S., Ortego, J., Pena, C.: Efficient human hand kinematics for manipulation tasks. In: IROS (2008)
11. Cordella, F., Zollo, L., Guglielmelli, E., Siciliano, B.: A bio-inspired grasp optimization algorithm for an anthropomorphic robotic hand. International Journal on Interactive Design and Manufacturing **6**(2), 113–122 (2012)
12. Dibra, E., Wolf, T., Oztireli, C., Gross, M.: How to refine 3d hand pose estimation from unlabelled depth data? In: 3DV (2017)
13. Drover, D., Chen, C.H., Agrawal, A., Tyagi, A., Phuoc Huynh, C.: Can 3d pose be learned from 2d projections alone? In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
14. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., Yuan, J.: 3d hand shape and pose estimation from a single RGB image. In: CVPR (2019)
15. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR (2020)
16. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR (2019)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
18. Heap, T., Hogg, D.: Towards 3D hand tracking using a deformable model. In: FG (1996)
19. Iqbal, U., Molchanov, P., Breuel, T., Gall, J., Kautz, J.: Hand pose estimation via latent 2.5d heatmap regression. In: ECCV (2018)
20. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)

21. Kuch, J.J., Huang, T.S.: Vision based hand modeling and tracking for virtual teleconferencing and telecollaboration. In: CVPR (1995)
22. Kulon, D., Wang, H., Güler, R.A., Bronstein, M., Zafeiriou, S.: Single image 3d hand reconstruction with mesh convolutions. In: BMVC (2019)
23. Lee, J., Kunii, T.L.: Model-based analysis of hand posture. IEEE Computer Graphics and applications **15**(5), 77–86 (1995)
24. Lin, J., Wu, Y., Huang, T.S.: Modeling the constraints of human hand motion. In: IEEE Workshop on Human Motion (2000)
25. Melax, S., Keselman, L., Orsten, S.: Dynamics based 3d skeletal hand tracking. In: ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (2013)
26. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: Ganerated hands for real-time 3d hand tracking from monocular RGB. In: CVPR (2018)
27. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: ICCV (2011)
28. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Efficient model-based 3d tracking of hand articulations using kinect. In: BMVC (2011)
29. Panteleris, P., Oikonomidis, I., Argyros, A.: Using a single rgb frame for real time 3d hand pose estimation in the wild. In: WACV (2017)
30. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
31. Reed, N.: What is the simplest way to compute principal curvature for a mesh triangle? `https://computergraphics.stackexchange.com/questions/1718/what-is-the-simplest-way-to-compute-principal-curvature-for-a-mesh-triangle` (2019)
32. Rhee, T., Neumann, U., Lewis, J.P.: Human hand modeling from surface anatomy. In: ACM SIGGRAPH symposium on interactive 3D graphics and games (2006)
33. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. In: SIGGRAPH-Asia (2017)
34. Ryf, C., Weymann, A.: The neutral zero method—a principle of measuring joint function. Injury **26**, 1–11 (1995)
35. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1145–1153 (2017)
36. Spurr, A., Song, J., Park, S., Hilliges, O.: Cross-modal deep variational hand pose estimation. In: CVPR (2018)
37. Sridhar, S., Mueller, F., Zollhoefer, M., Casas, D., Oulasvirta, A., Theobalt, C.: Real-time joint tracking of a hand manipulating an object from RGB-D input. In: ECCV (2016)
38. Sridhar, S., Oulasvirta, A., Theobalt, C.: Interactive markerless articulated hand motion tracking using RGB and depth data. In: ICCV (2013)
39. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: ICCV (2017)
40. Tekin, B., Bogo, F., Pollefeys, M.: H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In: CVPR (2019)
41. Tompson, J., Stein, M., Lecun, Y., Perlin, K.: Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics (ToG) **33**(5) (2014)
42. Wan, C., Probst, T., Gool, L.V., Yao, A.: Self-supervised 3d hand pose estimation through training by fitting. In: CVPR (2019)

43. Wu, Y., Huang, T.S.: Capturing articulated human hand motion: A divide-and-conquer approach. In: ICCV (1999)
44. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: CVPR (2019)
45. Xu, C., Cheng, L.: Efficient hand pose estimation from a single depth image. In: ICCV (2013)
46. Yang, L., Yao, A.: Disentangling latent hands for image synthesis and pose estimation. In: CVPR (2019)
47. Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., Yang, Q.: 3d hand pose tracking and estimation using stereo matching. arXiv:1610.07214 (2016)
48. Zhang, X., Li, Q., Mo, H., Zhang, W., Zheng, W.: End-to-end hand mesh recovery from a monocular RGB image. In: ICCV (2019)
49. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: ICCV (2017)
50. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: ICCV (2017)
51. 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: ICCV (2019)

Supplementary: Weakly Supervised 3D Hand Pose Estimation via Biomechanical Constraints

Here we provide additional implementation details and more experimental comparisons. In Section 1 we describe details on how the angle loss is computed and the joint angle interdependence is modeled. Section 2 repeats the ablation study on additional datasets (HO-3D) to highlight the generalizability of results. Section 3 demonstrates the effect of weak-supervision in two additional settings, one using a real dataset as the fully-supervised data and the other using MPII in-the-wild data as weak-supervision. Section 4 compares BMC to an adversarial loss. Sections 5 and 6 provide additional results of bootstrapping via weak-supervision with synthetic or real data. Section 7 shows further qualitative results of using BMC. Sections 8 and 9 provide additional implementation details and results on HANDS2019 challenge, respectively.

1 Joint angle loss

Joint angle ambiguity. The computation of the joint angles lead to ambiguities. More specifically, two different vectors on the unit sphere may map to the same joint angles.

For example, given two bones $\mathbf{b}_i^{\mathbf{F}_i,1} = [1,0,1]$ and $\mathbf{b}_i^{\mathbf{F}_i,2} = [-1,0,1]$ in a coordinate frame $\mathbf{F}_i$, we have using $\mathrm{P}_{xz}(\mathbf{b}_i^{\mathbf{F}_i,1}) = [1,0,1]$ and $\mathrm{P}_{xz}(\mathbf{b}_i^{\mathbf{F}_i,2}) = [-1,0,1]$:

$$
\begin{aligned}
\theta_i^{\mathrm{f},1} &= \alpha(\mathrm{P}_{xz}(\mathbf{b}_i^{\mathbf{F}_i,1}), \mathbf{z}_i) = \alpha([1,0,1], \mathbf{z}_i) \\
&= \pi/4 \\
\theta_i^{\mathrm{a},1} &= \alpha(\mathrm{P}_{xz}(\mathbf{b}_i^{\mathbf{F}_i,1}), \mathbf{b}_i^{\mathbf{F}_i,1}) \\
&= \alpha([1,0,1],[1,0,1]) = 0 \\
\theta_i^{\mathrm{f},2} &= \alpha(\mathrm{P}_{xz}(\mathbf{b}_i^{\mathbf{F}_i,2}), \mathbf{z}_i) = \alpha([-1,0,1], \mathbf{z}_i) \\
&= \pi/4 \\
\theta_i^{\mathrm{a},2} &= \alpha(\mathrm{P}_{xz}(\mathbf{b}_i^{\mathbf{F}_i,2}), \mathbf{b}_i^{\mathbf{F}_i,2}) \\
&= \alpha([-1,0,1],[-1,0,1]) = 0
\end{aligned}
\tag{1}
$$

Therefore, both bones map to the same angle pair $(\pi/4, 0)$. To resolve this, we perform an octant look up. Given the flexion angle θ_i^{f} and abduction angle θ_i^{a} of bone i, we negate the respective angle if the bone lies within the negative

x-octant or negative y-octant:

$$\theta_i^{\mathrm{f}} = \begin{cases} -\theta_i^{\mathrm{f}}, & \text{if } b_{i,x}^{\mathbf{F}_i} < 0 \\ \theta_i^{\mathrm{f}}, & \text{else} \end{cases} \\ \theta_i^{\mathrm{a}} = \begin{cases} -\theta_i^{\mathrm{a}}, & \text{if } b_{i,y}^{\mathbf{F}_i} < 0 \\ \theta_i^{\mathrm{a}}, & \text{else} \end{cases} \tag{2}$$

Where $b_{i,x}^{\mathbf{F}_i}$,$b_{i,y}^{\mathbf{F}_i}$ is the x/y-component of the bone vector given in coordinates of its local coordinate frame $\mathbf{F}_i$. This leads to angles in the range $\theta_i^{\mathrm{f}} \in [-\pi, \pi]$ and $\theta_i^{\mathrm{a}} \in [-\pi/2, \pi/2]$ respectively.

Approximation of Convex Hull. Fig. 1 plots the distribution of the pinkys MCP flexion/extension angles of the FH dataset, visualized as red points. The red rectangle corresponds to the valid range of angles when considering both angle limits independently. Hence the corners correspond to $(\min_i^f, \min_i^a), (\min_i^f, \max_i^a), (\max_i^f, \max_i^f), (\max_i^f, \min_i^a)$ in counter-clockwise order, where $\min_i^k, \max_i^k$ corresponds to the minimum/maximum of angle θ_i^k, where $k \in \{a, f\}$.

In order to take the dependence of the angle limits in account, we first compute the convex hull of the angle points. However, depending on the shape of the point cloud, the number of points lying on the hull can vary and be numerous. In order to keep the number of hull points low and consistent for all joint angles, we approximate this hull in two steps. We first employ the Ramer-Douglas-Peucker algorithm, a polygon simplification algorithm. This significantly reduces the number of vertices in the hull, but still results in a variable number. To ensure consistency, we apply a greedy algorithm that iteratively removes points such that the hull encompasses as many points as possible until we reach the desired number of points, resulting in our approximation $\mathcal{H}_i$. For all our experiments, we set number of points to be 10. The green polygon in Fig. 1 displays this approximation to the convex hull.

Distance computation. To compute the distance $\mathcal{H}_i$, we compute two values. The first indicates if an angle point $\boldsymbol{\theta}_i$ is contained within the hull. The second corresponds to the distance to the hull. Here we detail how we compute both values. For ease of notation, we assume that the points in $\mathcal{H}_i$ are ordered counter-clockwise beginning from any point in $\mathcal{H}_i$. Let $\mathcal{H}_{i,k}$ be the k-th point in $\mathcal{H}_i$. An edge $\mathbf{v}_k$ of the hull is given as:

$$\mathbf{v}_k = \mathcal{H}_{i,k+1} - \mathcal{H}_{i,k}, \text{ for } k \in [1, 10] \\ \mathbf{w}_k = \boldsymbol{\theta}_i - \mathcal{H}_{i,k}, \text{ for } k \in [1, 10] \tag{3}$$

Where we define $\mathcal{H}_{i,11} = \mathcal{H}_{i,1}$ to wrap around the hull.

To compute if a point $\boldsymbol{\theta}_i$ is contained within $\mathcal{H}_i$, we exploit the convexity of the hull and make use of the cross-product. Specifically, we compute the 2D cross-product between $\mathbf{v}_k$ and $\mathbf{w}_k$. Intuitively, if the cross-product $\mathbf{w}_k \times \mathbf{v}_k$ is positive for any given edge k, then the angle point lies outside of the hull. If its negative for all, it is contained within. If it lies on the hull, we consider it to be

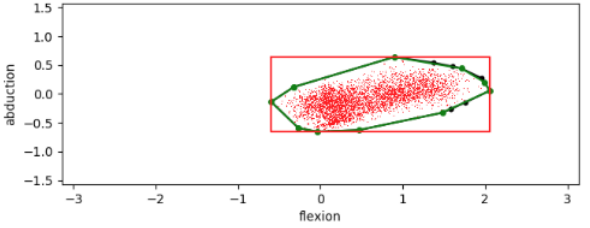


Fig. 1: $(\theta^{\mathrm{f}}, \theta^{\mathrm{a}})$-plane. Green: $\mathcal{H}_i$. Red: min/max-box

contained within it. More formally:

$$c = \prod_{k=1}^{10} \mathbb{1}_{(\mathbf{w}_k \times \mathbf{v}_k) \leq 0} \tag{4}$$

To compute the distance of $\boldsymbol{\theta}_i$ to the hull, we compute its distance to each edge and take the minimum. Given edge $\mathbf{v}_k$ and point $\boldsymbol{\theta}_i$, their distance is the minimum distance between either endpoints of $\mathbf{v}_k$ or the projection of $\mathbf{w}_k$ onto $\mathbf{v}_k$. Formally:

$$\begin{aligned} t &= \max(0, \min(1, \mathbf{w}_k^T \mathbf{v}_k / ||\mathbf{v}_k||_2^2)) \\ \mathbf{p}_k &= \mathcal{H}_{i,k} + t\mathbf{v}_k \\ D(\mathbf{v}_k, \boldsymbol{\theta}_i) &= |\cos(\boldsymbol{\theta}_i) - \cos(\mathbf{p}_k)| + |\sin(\boldsymbol{\theta}_i) - \sin(\mathbf{p}_k)| \end{aligned} \tag{5}$$

Where the min/max ensures that we do not extend beyond the endpoints of $\mathbf{v}_k$. Given the distance to the edge, we can compute the distance to the hull $\mathcal{H}_i$:

$$D(\boldsymbol{\theta}_i, \mathcal{H}_i) = \min_k D(\mathbf{v}_k, \boldsymbol{\theta}_i) \tag{6}$$

This formulation computes the distance towards $\mathcal{H}_i$, whether the point is contained or not. We do not want to penalize points that lie within the hull, as that constitutes our range of valid angles. Therefore we make use of the quantity c computed in Eq. 4, which leads to the final angle loss function:

$$D_A(\boldsymbol{\theta}_i, \mathcal{H}_i) = (1 - \mathbb{1}_c) D(\boldsymbol{\theta}_i, \mathcal{H}_i) \tag{7}$$

This returns a loss of 0 if the angle point $\boldsymbol{\theta}_i$ is contained, otherwise it returns the distance to the approximation of the convex hull $\mathcal{H}_i$. This constitutes our angle loss for bone i.

2 Ablation study

We repeat the ablation study with the HO-3D dataset. All evaluations are done on a custom split, where we manually extract two sequences for the test and use

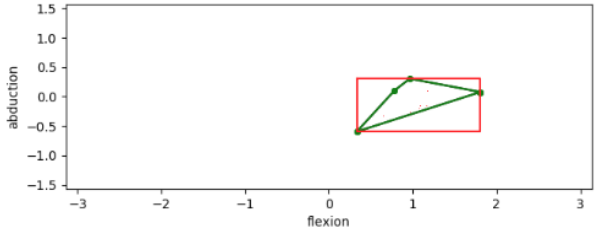


Fig. 2: $(\theta^{\mathrm{f}}, \theta^{\mathrm{a}})$-plane. Green: $\mathcal{H}_i$. Red: min/max-box

the remainder for the training set. Each error is computed for the root relative case.

Refinement network. We train two models using full supervision on HO-3D ($\mathbf{3D}_{\mathrm{HO3D}}$). The first model (w/o refinement) does not use the proposed refinement network, whereas the second does (w.refinement). We showcase the performance difference in the first row of Tab. 1. We note a reduction of 2.97mm mean error when using the refinement network.

BMC ablation. We study the individual contribution of the BMC losses. We bootstrap the 3D annotation from synthetic data and use only the 2D annotation of HO-3D. The first model constitutes our baseline, which is trained only on that data ($\mathbf{3D}_{\mathrm{RHD}} + \mathbf{2D}_{\mathrm{HO3D}}$). We incrementally add the bone length loss $\mathcal{L}_{\mathrm{BL}}$, the root bone loss $\mathcal{L}_{\mathrm{RB}}$ and lastly the angle loss $\mathcal{L}_{\mathrm{A}}$. We train a fully supervised model ($\mathbf{3D}_{\mathrm{RHD}} + \mathbf{3D}_{\mathrm{HO3D}}$) which is our upper bound. We refer to the second section of Tab. 1. Each loss contributes towards a reduction in mean error, culminating in a total decrease of 5.21mm as compared to our 2D only baseline.

Co-dependency between angles. We train two models. The first models the angle limits independently, whereas the second takes the dependency of the limits into account. The resulting performance is shown in Tab. 1. We note a minor performance degradation. We attribute this to the extremely limited angle range contained in the HO-3D dataset. As it contains subjects holding various object in a gripping pose while rotating it in front of the camera, the actual angles of the fingers do not change. Therefore the range of angles across the dataset is low, which leads to a very tight angle limit. This does not generalize well, which in turn hurts performance. Fig. 2 displays the angle-plane plot for HO-3D using the pinkys MCP flexion/extension angles. Comparing with Fig. 1, which plots the plane for the same finger for FH, we see that the resulting range of HO-3D is a lot more severely limited. This is to be expected, as HO-3D is a very constrained dataset due to the aforementioned reason.

BMC limits. We study the effect of approximating the BMC limits when using a different dataset to compute these values. We compute the hand parameters from RHD and perform the same weakly-supervised experiment as previously ($\mathbf{3D}_{\mathrm{RHD}} + \mathbf{2D}_{\mathrm{HO3D}}$). As can be seen in the last row of Tab. 1, we note a slight increase in loss, however it still clearly outperforms the 2D baseline in mean error (18.50 mm vs 23.71 mm).

Table 1: Ablation studies on **validation** split of HO-3D. The models of the first section was trained on our train split of HO-3D.

HO-3D	3D Pose Estimation (root-relative) EPE (mm) mean ↓	median ↓	AUC ↑
Effect of Z^{root} refinement			
w/o refinement	25.34	24.39	0.79
w. refinement	22.37	23.01	0.83
Effect of BMC components			
$\mathbf{3D}_{\mathrm{RHD}} + \mathbf{2D}_{\mathrm{HO3D}}$	23.71	22.07	0.78
$+ \mathcal{L}_{\mathrm{BL}}$	22.15	20.27	0.80
$+ \mathcal{L}_{\mathrm{RB}}$	18.83	17.79	0.87
$+ \mathcal{L}_{\mathrm{A}}$	18.50	17.41	0.87
$\mathbf{3D}_{\mathrm{RHD}} + \mathbf{3D}_{\mathrm{HO3D}}$	16.74	16.94	0.89
Effect of angle co-dependency			
Independent	18.30	17.40	0.87
Dependent	18.50	17.41	0.87
Effect of BMC limits			
Approximated	19.21	17.88	0.86
Computed	18.50	17.41	0.87

3 Effect of Weak-Supervision

We repeat the experiments of Section 5.3 in the main paper using different datasets. We show that the effect of weak-supervision also holds when using fully labeled real data or weakly-labeled in-the-wild data.

STB. We reproduce the results of Section 5.3 in the main paper, but instead of using RHD we use STB [47] as the fully supervised dataset. The weakly-supervised dataset remains FH. The purpose of this experiment is to demonstrate that the effect of weak supervision also takes place when using a real dataset for full supervision. Table 2 (top) shows the result.

MPII - in-the-wild dataset. We reproduce the results of Section 5.3 in the main paper, but using MPII [35] as our weakly-supervised dataset. This is to demonstrate the effect of weak-supervision stemming from datasets collected in-the-wild, a potentially useful supervision source. We evaluate on the validation split of FH. Table 2 (bottom) shows the result. Note that as the MPII dataset only contains 2D labels and no 3D annotation is provided, the fully supervised upper bound cannot be performed and is therefor omitted from the table.

4 Comparison with Adversarial loss

It is intuitive to think of drawing parallels between BMC and an adversarial loss. BMC can be interpreted as a discriminator penalizing poses that do not adhere to the distribution of valid hand poses. However, BMC models the task at hand more closely and only requires the limits, whereas a discriminator requires access to a full dataset of 3D poses. In order to see how a discriminator performs against

Table 2: This table show-cases the same effect of weak-supervision as Table 2 in the main paper but evaluated in different settings. All models are evaluated on the **validation** split of FH. (top) We use STB as the fully labeled dataset and supplement is using weakly-labeled FH. (bottom) We use RHD as the fully labeled dataset and MPII as the weakly-supervised data. The same trend can be observed in both settings. Adding weakly-supervised data improves 3D prediction performance due to predicted 3D poses with the correct 2D projection. By incorporating our proposed biomechanically constraints we significantly improve 3D pose accuracy due to more accurate **Z**. Note that as the MPII dataset only contains 2D labels and no 3D annotation is provided, the fully supervised upper bound cannot be performed and is therefor omitted from the table.

Effect of weak-supervision	Description	mean ↓ 2D (pixel)	mean ↓ Z (mm)	mean ↓ 3D (mm)
	3D labels: STB			
$\mathbf{3D}_{\text{STB}} + \mathbf{3D}_{\text{FH}}$	Fully supervised, real	3.85	5.68	9.05
+ $\mathcal{L}_{\mathbf{BMC}}$ (**ours**)	+ BMC	**3.83**	**5.50**	**8.89**
$\mathbf{3D}_{\text{STB}}$	Fully sup. lower bound	20.45	36.80	54.92
+ $\mathbf{2D}_{\text{FH}}$	+ Weakly supervised, real	**3.86**	35.41	42.02
+ $\mathcal{L}_{\mathbf{BMC}}$ (**ours**)	+ BMC	3.88	**11.17**	**18.58**
	2D labels: MPII			
$\mathbf{3D}_{\text{RHD}}$	Fully supervised, synthetic only	12.35	20.02	30.82
+ $\mathbf{2D}_{\text{MPII}}$	+ Weakly supervised, real	10.36	19.77	28.81
+ $\mathcal{L}_{\mathbf{BMC}}$ (**ours**)	+ BMC	**10.35**	**17.72**	**27.10**

BMC, we perform an experiment in the same setting as the ablation study. We train on fully supervised RHD and weakly-supervised FH, and evaluate on the validation split of FH. As it has not been shown if and how the adversarial loss works for the task of 3D hand pose estimation, we adapt a model from literature applied to 2D body pose [13]. In order to adjust to the new setting, we performed a search for the optimal hyperparameters to improve the performance of the discriminator. We show the results in Table 3. As can be seen, BMC outperforms the adversarial loss. We hypothesise this is due to BMC modeling the task at hand more closely.

5 Bootstrapping with Synthetic Data

We show the full results of the online evaluation on FH and HO-3D in Table 4.

6 Bootstrapping with Real Data

Tab. 5 shows the full result of Bootstrapping with real data, as evaluated by the online submission system [1]. Recall that we assume the remainder of the data to be weakly-supervised, i.e it contains the 2D annotation. We list the exact number

[1] `https://competitions.codalab.org/competitions/21238`

Table 3: We compare using BMC to an adversarial loss adapted from [13]. BMC outperforms the adversarial loss. We hypothesise this is due to BMC modeling the task at hand more closely.

Comparison to adversarial loss	Description	3D Pose Estimation (root-relative) EPE (mm) mean ↓	median ↓	AUC ↑
$\mathbf{3D}_{\mathrm{RHD}} + \mathbf{2D}_{\mathrm{FH}}$	Baseline	20.92	16.93	0.81
$\mathbf{3D}_{\mathrm{RHD}} + \mathbf{2D}_{\mathrm{FH}} + \mathcal{L}_{\mathbf{BMC}}$	BMC	**15.48**	**13.49**	**0.91**
$\mathbf{3D}_{\mathrm{RHD}} + \mathbf{2D}_{\mathrm{FH}} + \mathcal{L}_{\mathbf{adv}}$	Adversarial	17.60	14.38	0.87

Table 4: Bootstrapping results on the respective **test split**, as evaluated by the *online submission system*. Results are given in mm.

FH	Description	aligned mean ↓	aligned AUC ↑	unaligned mean ↓	unaligned AUC ↑
Zimmermann et al. [51]	fully supervised FH	1.10	0.78	**7.13**	0.19
$\mathbf{3D}_{\mathrm{RHD}} + \mathbf{3D}_{\mathrm{FH}}$	fully supervised RHD/FH	**0.90**	**0.82**	7.54	**0.20**
$\mathbf{3D}_{\mathrm{RHD}}$	fully supervised RHD	1.60	0.69	15.15	0.06
$+ \mathbf{2D}_{\mathrm{FH}}$	+ weakly-supervised FH	1.26	0.75	13.02	0.14
$+ \mathcal{L}_{\mathbf{BMC}}$	+ BMC	**1.13**	**0.78**	**10.39**	**0.15**
HO3D	Description	EXTRAP ↓	INTERP ↓	OBJECT ↓	SHAPE ↓
$\mathbf{3D}_{\mathrm{RHD}} + \mathbf{3D}_{\mathrm{HO3D}}$	fully supervised HO3D	18.22	5.02	16.56	10.79
$\mathbf{3D}_{\mathrm{RHD}}$	fully supervised RHD	20.84	33.57	35.08	23.94
$+ \mathbf{2D}_{\mathrm{HO3D}}$	+ weakly supervised HO3D	19.57	25.16	25.79	21.05
$+ \mathcal{L}_{\mathbf{BMC}}$	+ BMC	**18.42**	**10.31**	**19.91**	**12.51**

of 3D labeled samples used, in addition to the percentage wrt. to the entire dataset it corresponds to. Note that the percentage values have been rounded for readability, but the number of samples is exact. We divide the table according to three categories a) **Aligned / Unaligned** - Procrustes analysis is used to align before computing the score b) **Mean / AUC** - The AUC is given for PCK values that lie in an interval from 0 mm to 50 mm with 100 equally spaced thresholds. c) **With / Without BMC** - Using our proposed biomechanical constraints. We first focus on the aligned results. Using BMC, the required amount of 3D annotated data for a given AUC is approximately *halved*. This trend continues for labeling percentages up to $\sim 13\%$. For example, to achieve the same performance as a model that is trained without BMC on 3810 3D labeled data samples, BMC achieves the same performance with 1993 3D labeled samples, roughly half the amount.

A similar trend can be observed for the unaligned score. For labeling percentages up to 6.8% (1993), the required amount of data to reach the same performance is approximately halved (997).

7 Qualitative results

We show qualitative results of the Bootstrapping with Synthetic Data experiment in Fig. 3. We display the predicted $\mathbf{J}^{3D}$ of both $\mathbf{3D}_{\mathrm{RHD}} + \mathbf{2D}_{\mathrm{FH}}$ (w/o BMC) and

Table 5: Scores as evaluated on the online submission system. The first column denotes the percentage (in brackets) of 3D annotated samples used during training, where the remainder is annotated only with 2D labels. Note that the percentages are rounded, but the number of samples are exact. + **indicates the model trained with BMC,** − **indicates the model trained without it.**

FH	+: with BMC (**ours**)	Aligned				Unaligned			
	−: without BMC	mean ↓		AUC ↑		mean ↓		AUC ↑	
3D samples: Number	3D samples: Perc.	−	+	−	+	−	+	−	+
1	(3.4*e*–3%)	1.96	1.64	0.62	0.68	34.86	18.40	0.08	0.11
5	(0.017%)	1.85	1.41	0.64	0.72	26.40	15.26	0.11	0.13
14	(0.045%)	1.78	1.39	0.65	0.73	25.24	12.98	0.11	0.13
27	(0.094%)	1.75	1.34	0.66	0.73	23.90	11.93	0.12	0.14
127	(0.43%)	1.54	1.24	0.70	0.76	21.83	12.08	0.13	0.16
499	(1.7%)	1.23	1.18	0.76	0.77	11.68	10.88	0.17	0.18
997	(3.4%)	1.14	1.12	0.77	0.78	9.85	9.42	0.18	0.19
1993	(6.8%)	1.10	1.07	0.78	0.79	8.83	8.75	0.19	0.20
3810	(13%)	1.06	1.04	0.79	0.79	8.01	7.90	0.21	0.21
7327	(25%)	1.02	1.01	0.80	0.80	7.91	7.84	0.21	0.21
14653	(50%)	0.99	1.00	0.80	0.80	7.46	7.56	0.22	0.22
29305	(100%)	0.98	0.98	0.81	0.81	7.18	7.18	0.23	0.23

$\mathbf{3D}_{\text{RHD}} + \mathbf{2D}_{\text{FH}} + \mathcal{L}_{\text{BMC}}$ (w. BMC). Two views are shown. The first displays the view from the front or camera view (looking in direction of the z-axis), the second shows the view from the top of the world space, looking down (looking in the *opposite* direction of the x-axis). Additionally, we plot the 2D predictions of both models, where green corresponds to without BMC and red is the model using BMC.

We see that despite both models predicting accurately the 2D pose, its predicted 3D pose are different. Not using BMC, the model predicts bio-physically implausible poses. This is due to unseen 3D poses, views and occlusions. Additionally, the 3D component of the model has only been trained on synthetic data. For example, RHD does not contain object occlusions or ego-centric views. Using BMC, our model can better adapt its depth-component during training to these unseen 3D poses, resulting in more accurate predictions.

8 Architecture and training

We use a standard ResNet-50 network for our backbone. We replace the last linear layer to output a 21×3 dimensional vector. The first two dimensions correspond to the $2D$ keypoints, whereas the last layer corresponds to the root-relative depth Z^{r}.

Our Z^{root} refiner consists of a three layered MLP, using leaky ReLU non-linearity. We used BatchNorm in between all layers except the last. For the

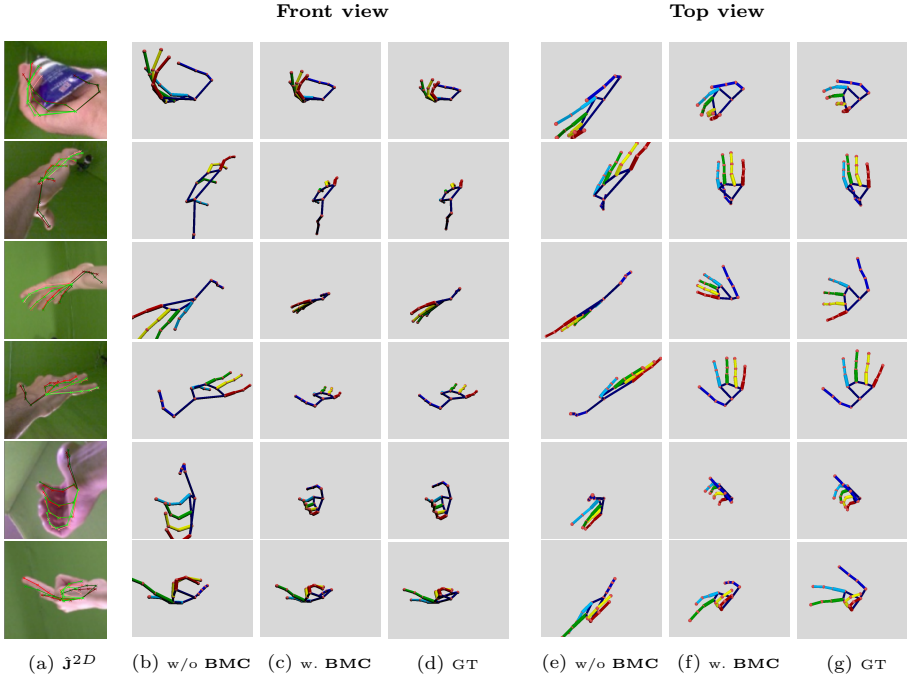


Fig. 3: Qualitative results of the Boostrapping with Synthetic Data experiment. Testing performed on custom split of FH. Fig. 3a: We see that the model trained without BMC (green), as well as the model trained with BMC (red), perform equally well on the 2D prediction task. Fig. 3d, Fig. 3g show the ground-truth joint skeleton from the camera view, as well as the "top" view looking down, respectively. Fig. 3b and Fig. 3e show the 3D predictions of the model trained fully supervised on RHD and weakly-supervised on FH. Despite the accurate 2D predictions, the 3D pose is incorrect, displaying implausible bio-physical poses. Fig. 3c and Fig. 3f show the result of incorporating BMC into the model. The predictions are kinematically and structurally sound, and as a result closer to the ground-truth predictions.

Table 6: Architecture of the refinement network. It takes the predicted and calculated values $\mathbf{z}^r \in \mathbb{R}^{21}$, $\mathbf{K}^{-1}\mathbf{J}^{2D} \in \mathbb{R}^{21\times 3}$, $Z^{root} \in \mathbb{R}$ and outputs a residual term r such that $\hat{Z}^{root}_{\text{ref}} = \hat{Z}^{root} + r$

Refinement Network
Linear(85, 128)
LeakyReLU(0.01)
BatchNorm
Linear(128, 128)
LeakyReLU(0.01)
BatchNorm
Linear(128, 1)

Table 7: HANDS2019 challenge results on the **test split** of HO-3D, as evaluated by the *online submission system*. All methods were trained only on HO-3D. We show the top four submission. The winner was selected based on the extrapolation score. Results are given in mm.

HO-3D	EXTRAP ↓	INTERP ↓	OBJECT ↓	SHAPE ↓
Ours	**24.74**	6.70	27.36	**13.21**
Nplwe	29.19	**4.06**	**18.39**	15.79
lin84	31.51	19.15	30.59	23.47
Hasson et al. [16]	38.42	7.38	31.82	15.61

cross-dataset evaluation, we empirically found that not using BatchNorm resulted in better accuracy. The exact architecture using BatchNorm is listed in Tab. 6.

The network was trained for 70 epochs using SGD with a learning rate of $5e$–3 and a step-wise learning rate decay of 0.1 after every 30 epochs.

We set the weight values as follows: $\lambda_{2D} = 1$, $\lambda_{Z^{\text{r}}} = 5$, $\lambda_{Z^{\text{root}}} = 1$. For all experiments using BMC, we set the individual weights of the losses as follows: $\lambda_{\text{BL}} = 0.1$, $\lambda_{\text{RB}} = 0.1$, $\lambda_{\text{A}} = 0.01$

9 HANDS2019 challenge

The HANDS2019 challenge[2] was organized to evaluate cutting edge methods for 3D hand pose estimation. The rules of challenge task #3 required us to train solely on the HO-3D dataset. We trained the proposed model without auxiliary losses. The refinement step was vital for achieving the first place of the competition, demonstrating the performance of the underlying backbone model.

[2] `https://competitions.codalab.org/competitions/21116`