3D Hand Pose Estimation Using Synthetic Data and Weakly Labeled RGB Images

Yujun Cai, Liuhao Ge, Jianfei Cai, Nadia Magnenat Thalmann, Junsong Yuan

Abstract—Compared with depth-based 3D hand pose estimation, it is more challenging to infer 3D hand pose from monocular RGB images, due to the substantial depth ambiguity and the difficulty of obtaining fully-annotated training data. Different from the existing learning-based monocular RGB-input approaches that require accurate 3D annotations for training, we propose to leverage the depth images that can be easily obtained from commodity RGB-D cameras during training, while during testing we take only RGB inputs for 3D joint predictions. In this way, we alleviate the burden of the costly 3D annotations in real-world dataset. Particularly, we propose a weakly-supervised method, adaptating from fully-annotated synthetic dataset to weakly-labeled real-world single RGB dataset with the aid of a depth regularizer, which serves as weak supervision for 3D pose prediction. To further exploit the physical structure of 3D hand pose, we present a novel CVAE-based statistical framework to embed the pose-specific subspace from RGB images, which can then be used to infer the 3D hand joint locations. Extensive experiments on benchmark datasets validate that our proposed approach outperforms baselines and state-of-the-art methods, which proves the effectiveness of the proposed depth regularizer and the CVAE-based framework.

Index Terms—3D hand pose estimation, weakly-supervised methods, depth regularizer, pose-specific subspace.

1 Introduction

Hands are of central importance to humans, since they provide a natural way for humans to manipulate objects and communicate with each other. Articulated hand pose estimation has aroused a long-standing research attention in the past decades [1], [2], [3], [4], [5], [6], [7], [8], as it serves as an essential component in numerous applications such as human-computer interaction, virtual reality, robotics and rehabilitation. Additionally, similar to human body pose estimation used for action recognition [9], [10], [11], 3D hand pose estimation can be further applied to gesture recognition and sign language recognition [12], [13].

Many recent works [12], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24] on 3D hand pose estimation have gained tremendous success due to the availability of low-cost depth cameras and the large public 3D hand pose datasets [3], [4], [25], [26] with depth images. The advance in monocular RGB-based 3D hand pose estimation [27], [28], [29], [30], however, still remains limited. Due to low cost and low power of RGB cameras, RGB-based solutions for 3D hand pose estimation are more favored than depth-based solutions in many vision applications.

Compared with depth images, single-view RGB images exhibit inherent depth ambiguity and lighting-sensitive appearance, which makes 3D hand pose estimation from single RGB images a challenging problem. Most recent works on RGB-based 3D hand pose estimation heavily rely on large amount of labeled data for training, while comprehensive real-world dataset with complete 3D annotations is often difficult to obtain, thus limiting the performance. Specifically, compared with 2D annotations, providing 3D annotations for real-world RGB images is typically more difficult since 2D locations can be directly defined in the RGB images while 3D locations cannot be easily labeled by human annotators. To address this problem, Zimmermann *et al.* [27] turned to render low-cost synthetic hands with 3D models, from which the ground truth of 3D joints can be easily obtained. Although achieving good performance on the synthetic dataset, this method does not generalize well to real image dataset due to the domain gap between image features. To tackle this issue, Mueller *et al.* [28] leveraged CycleGANs [31] to generate a "real" dataset transferred from synthetic dataset and combined a CNN with a kinematic 3D hand model for 3D pose estimation. However, its limited performance shows that there still exists gap between generated "real" images and real-world images. Mahdi *et al.* [32] proposed a domain transfer method which learns the feature mapping from color images to depth images. Recently, Spurr *et al.* [29] employed a cross-modal deep variational hand pose estimation, which learns a cross-modal latent representation that are estimated from different modalities. This method, still relying on 3D annotations of real hands, is thus different from our weakly-supervised setting where no 3D annotations of real images are provided.

Most of the previous works [27], [28], [29] for hand pose estimation from real-world single-view RGB images focus on training with complete 3D annotations, which are expensive and time-consuming to obtain, while ignoring the depth images that can be easily captured by commodity RGB-D cameras. In addition, such depth images contain rich cues for 3D hand pose labels, as depth-based methods

- *Yujun Cai and Liuhao Ge are with the Institute for Media Innovation, Nanyang Technological University, Singapore 639798. E-mail: {yujun001, ge0001ao}@e.ntu.edu.sg*
- *Jianfei Cai is with Faculty of Information Technology, Monash University. Email: jianfei.cai@monash.edu*
- *Nadia Magnenat Thalmann is with Institute of Media Innovation, Nanyang Technological University. Email: nadiathalmann@ntu.edu.sg*
- *Junsong Yuan is with Computer Science and Engineering Department, University at Buffalo. Email: jsyuan@buffalo.edu*

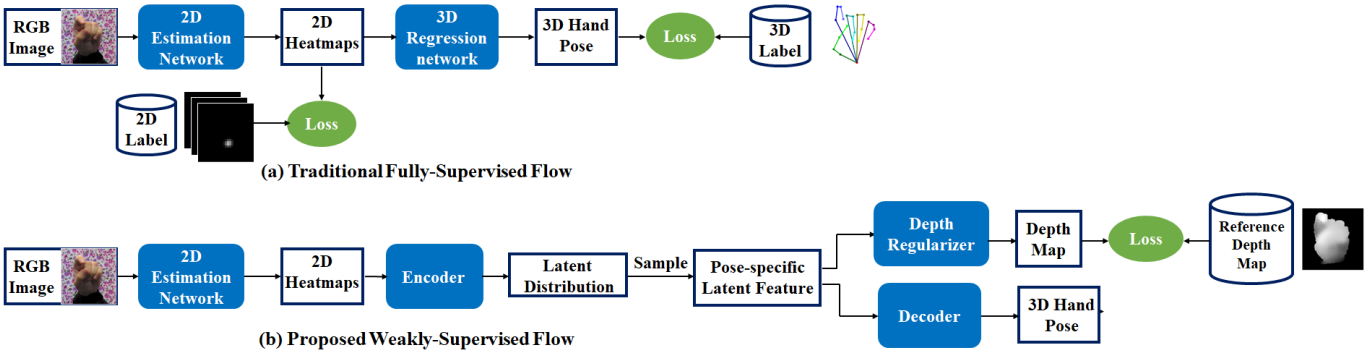


Fig. 1: Illustration of the concept of weakly supervised 3D hand pose estimation. Different from conventional fully-supervised methods (a) that use 3D labels to guide joint predictions, our proposed weakly-supervised method (b) leverages the reference depth map, which can be easily obtained by consumer-grade depth camera, to provide weak supervision. To better constrain the physical structure of 3D hand pose, we present a CVAE-based framework to learn the pose-specific latent distribution, from which we can sample the latent feature and further infer the 3D hand pose and render the corresponding depth map. Note that we only need the reference depth map during training as a regularizer. During testing, the trained model can predict 3D hand pose from RGB-only input.

achieve state-of-the-art performance on 3D pose estimation. Based on these observations, we propose to leverage the easily captured depth images to compensate the scarcity of entire 3D annotations during training, while during testing we take only RGB inputs for 3D hand pose estimation. Furthermore, to better constrain the physical structure of 3D hand pose, we deploy a statistical framework based on Conditional Variational Autoencoder (CVAE) to encode the pose-specific latent subspace, from which we can further decode the 3D hand joint locations. Fig. 1 illustrates the concept of our proposed weakly supervised 3D hand pose estimation method, which alleviates the burden of the costly 3D annotations in real-world datasets.

In particular, similar to the previous works [33], [34], [35], [36], [37] in body pose estimation, we split the task into 2D detection followed by a 2D-3D lifting step. Note that both synthetic and real datasets are utilized in our framework. For well-labeled synthetic data, different from most of the previous approaches that directly take a regression framework to map 2D to 3D output, we propose to learn a pose-specific latent distribution via an extension of CVAE framework, which aims to maximize the posterior probability of 3D hand pose estimation given the RGB image input. Specifically, we provide a framework that allows the training of a latent subspace encoded from the image feature domain and the 3D hand pose domain separately, and force the two embedded latent distributions close to each other. By doing this, the fully trained network model is able to capture the pose-specific latent subspace efficiently and robustly from the RGB images during testing, which can then be used to sample the pose-specific feature and infer the 3D hand joint locations.

For unlabeled real images, directly transferring the network trained on synthetic dataset to real-world dataset usually produces poor estimation accuracy, due to the domain gap between them. To address this problem, inspired by [38], [39], we innovate the structure with a depth regularizer, which generates depth images from the above mentioned pose-specific latent feature and regularizes the latent representation by supervising the generated depth map, as shown in Figure 1 (b). This network essentially learns the mapping from the pose-specific latent feature to its corresponding depth map, which can be used for the knowledge transfer from the fully-annotated synthetic images to unlabeled real-world images without entire 3D annotations. The effectiveness of the depth regularizer is experimentally verified by our weakly-supervised, semi-supervised and fully-supervised methods on benchmark datasets.

Compared with the existing methods for 3D hand pose estimation from monocular RGB images, our main contributions are:

- We innovatively introduce the weakly supervised problem of leveraging low-cost depth maps during training for 3D hand pose estimation from RGB images, which alleviates the burden of 3D joint labeling. To the best of our knowledge, we are the first ones to introduce such weakly-supervised setting for 3D hand pose estimation from RGB images.
- We propose an end-to-end learning based solution for weakly-supervised adaptation from fully-annotated synthetic images to real-world images without entire 3D annotations. Particularly, we introduce a depth regularizer supervised by the easily captured depth images, which considerably enhances the estimation accuracy compared with other baselines.
- We conduct experiments on the two benchmark datasets, which show that our weakly-supervised approach compares favorably with existing works and our proposed semi-supervised and fully-supervised methods are superior to the state-of-the-art methods.

This paper is an extension of our conference paper [40]. The new contributions of this paper include:

- We replace the original regression network with a novel CVAE-based statistical framework, which encodes the distribution of the pose-specific latent sub-

space that can be sampled to infer the 3D hand joint locations. Experimental results show that, the CVAE-based framework is able to improve the performance of prediction accuracy and provide a physically-valid structure for 3D hand pose compared with our original regression network.
- To further alleviate the burden of 2D annotation, we investigate our weakly-supervised method trained with only depth regularization, which removes the 2D supervision in the original weakly-supervised approach trained with both 2D labels and depth regularizer, as presented in our conference paper. Moreover, experimental results show that our proposed weakly-supervised method trained with only depth regularizer performs comparably with the original version constrained by both 2D labels and depth regularization.
- We modify the depth regularizer with the input of low-dimensional pose-specific feature instead of the predicted 3D hand pose. Moreover, we experimentally validate that the network can benefit more from constraining the latent representation by the proposed depth regularizer.
- We conduct more extensive experiments including in the semi-supervised setting and self-comparisons. Additionally, we also compare with more state-of-the-art methods on RHD [27], STB [41], Dexter Object [42] and Egodexter [43] datasets. Experimental results show that our method can achieve good performance and has plausible generalization ability in all settings.

The remainder of this paper is organized as follows: We first discuss the related work in Section 2 and then introduce our detailed methodology in Section 3. After that, Section 4 provides the experimental evaluations and Section 5 concludes this paper.

2 Related Work

3D hand pose estimation Articulated 3D hand pose estimation has been studied extensively for a long time, with vast theoretical innovations and important applications. Early work [1], [44], [45] on 3D hand pose estimation from monocular color input used complex model-fitting schemes which require strong prior knowledge on physics or dynamics and multiple hypotheses. These sophisticated methods, however, usually suffer from low estimation accuracy and restricted environments, which result in limited prospects in real-world applications. While multi-view approaches [46], [47] alleviate the occlusion problem and provide decent accuracy, they require sophisticated mesh models and optimization strategies that prohibit them from real-time tasks.

The emergence of low-cost consumer-grade depth sensors in the last few years greatly promotes the research on depth-based 3D hand pose estimation, since the captured depth images provide richer context that significantly reduces depth ambiguity. With the prevailing of deep learning technology [48], learning-based 3D hand pose estimation from single depth images has also been introduced, which can achieve state-of-the-art 3D pose estimation performance in real time. In general, they can be classified into generative approaches [49], [50], [51], discriminative approaches [12], [14], [17], [52], [53], [54], [55] and hybrid approaches [6], [56], [57], [58].

Inspired by the great improvement of CNN-based 3D hand pose estimation from depth images [59], deep learning has also been adopted in some recent works on monocular RGB-based applications [27], [28], [29], [60], [61], [62], [63]. In particular, Zimmermann and Brox [27] proposed a deep network that learns an implicit 3D articulation prior of joint locations in canonical coordinates, as well as constructing a synthetic dataset to tackle the problem of insufficient annotations. Spurr *et al.* [29] presented a "cross-modal variational model" based on VAE to learn a shared latent space between different modalities. Mueller *et al.* [28] embedded a "GANerated" network which transfers the synthetic images to "real" ones so as to reduce the domain shift between them. The performance gain achieved by these methods indicates a promising direction, although estimating 3D hand pose from single-view RGB images is far more challenging due to the absence of depth information. Our work, as a follow-up exploration, aims at alleviating the burden of 3D annotations in real-world dataset by bridging the gap between fully-annotated synthetic images and unlabeled real-world images.

For our proposed depth regularizer, Dibra *et al.* [39] is the closest work in spirit, which presented an end-to-end network that enables the adaptation from synthetic dataset to unlabeled real-world dataset. However, we want to emphasize that our method is significantly different from [39] in several aspects. Firstly, our work is targeted at 3D hand pose estimation from single RGB input, whereas [39] focuses on depth-based predictions. Secondly, compared with [39] that leverages a rigged 3D hand model to synthesize depth images from 3D hand pose, we use a simple convolutional network to infer the corresponding depth maps from the pose-specific latent feature. Our weakly-supervised adaptation is the first learning-based attempt that introduces a depth regularizer to monocular-RGB based 3D hand pose estimation. This presents a weakly-supervised solution for this problem and will enable further research of utilizing depth images in RGB-input applications.

Recovering 3D mesh from RGB images To further evaluate the deformable and articulated architecture of 3D hands, recently some deep learning based approaches [64], [65], [66], [67], [68], [69], [70] proposed to reconstruct 3D hand mesh from RGB inputs. For instance, [66], [67], [68], [69], [70] attempted to regress MANO parameters using deep neural networks with the supervisions of silhouette and/or 2D keypoints. Hasson *et al.* [64] presented a learnable model for reconstructing hand and object meshes during manipulations and exploited a novel contact loss that favors physically plausible hand-object constellations. Ge *et al.* [65] and Kulon *et al.* [71] leveraged graph-based methods to directly operate on the 3D mesh vertices and estimate 3D pose from the generated mesh. These approaches show a good direction on joint 3D pose and shape estimation.

Latent representation The concept of latent representation has been previously discussed in the literature [72], [73], which argues that the large degrees-of-freedom of 3D pose configurations are not independent with each other,

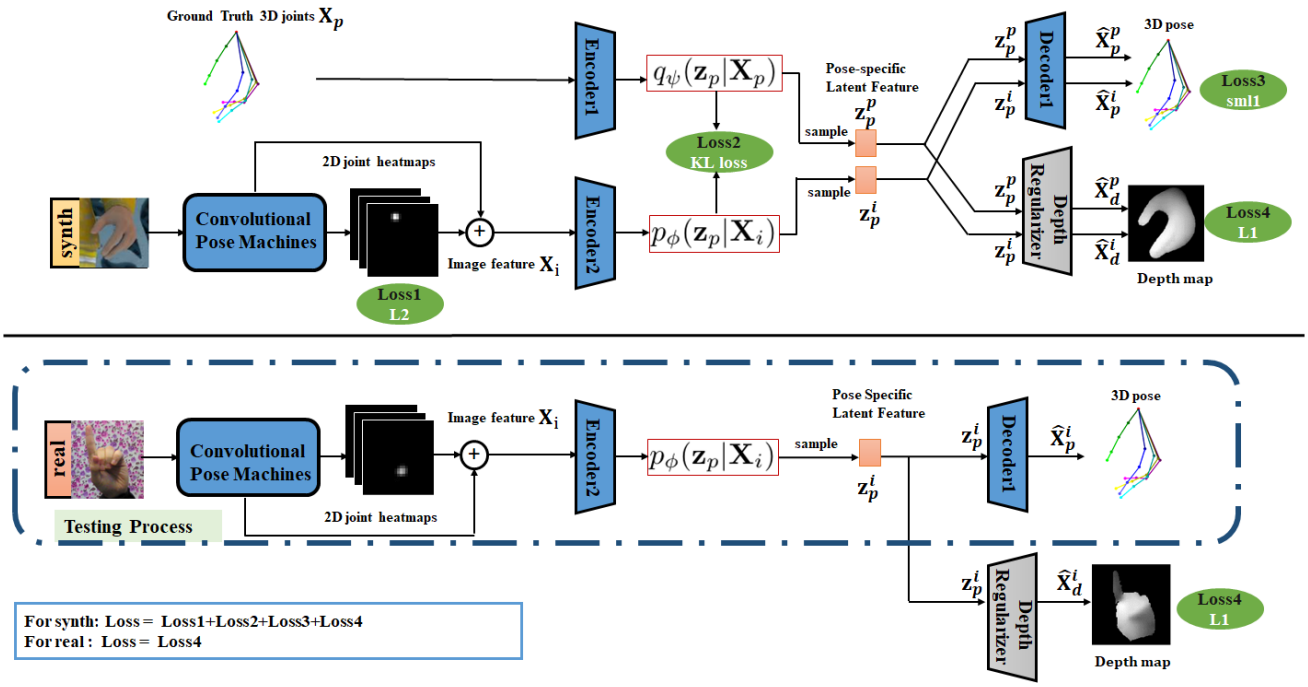


Fig. 2: Overview of our proposed weakly-supervised 3D hand pose estimation network, which is trained in an end-to-end manner. During training, cropped images from both synthetic dataset and real image dataset are mixed in each single batch as the input to the network. For well-labeled synthetic data, a pose specific low-dimensional latent feature is jointly learned from the 3D pose domain and the image feature domain, and can be further utilized to predict the 3D hand joint locations. To compensate the absence of ground truth annotations for unlabeled real data, we innovatively extend the network with a depth regularizer by leveraging the corresponding depth maps available in both synthetic and real datasets, so as to provide a weak supervision on the pose-specific latent embedding. During testing, real images only go through the part of the network in the dash-dotted line box. Note that modules with the same name share weights with each other.

and can be constrained in a low-dimensional subspace. For instance, [74], [75] applied the linear dimensionality reduction technique, PCA, to learn the pose subspace. Poier *et al.* [76] observed that the pose is predictive for the appearance of the hand seen from another view and tried to learn a pose specific representation using unlabeled data by constraining the different views of hand generated from the latent representation. Wan *et al.* [77] targeted at 3D hand pose estimation from depth images and proposed a network architecture based on two generative networks, including a variational auto encoder (VAE) for hand poses and a generative adversarial network (GAN) for modeling the distributions of depth images. To connect the separate latent spaces of depth images and hand pose, a one-to-one mapping function is introduced to align the two domains. Yang *et al.* [78] introduced a disentangled latent space to separate image variations such as image background content and hand pose, which can be used for image synthesis and 3D hand pose estimation tasks. Spurr *et al.* [29] learned a shared latent space that crosses multiple hand modalities such as depth and RGB images. Specifically, given a set of modalities, separate VAE networks are trained iteratively with one input modality contributing to the back-propagation per iteration.

Different from [29], [77], [78], our CVAE-based statistical framework builds on the extension of the conditional variational auto-encoder (CVAE) [79] and is trained to maximize the conditional log-likelihood of the predicted 3D hand pose based on the observation of RGB image input. We note that for fully-supervised and semi-supervised methods, our method allows for the joint learning of the pose-specific subspace from both 3D pose domain and image feature domain, as opposed to the iterative training strategy proposed in [29], while for weakly-supervised method, the pose-specific latent feature encoded from image features can be constrained by our proposed depth regularizer.

3 Methodology

3.1 Overview

Our target is to infer 3D hand pose from a monocular RGB image, where the 3D hand pose is represented by a set of 3D joint coordinates $\mathbf{\Phi} = \{\phi_k\}_{k=1}^{K} \in \mathbf{\Lambda}_{3D}$. Here $\mathbf{\Lambda}_{3D}$ is the $K \times 3$ dimensional hand joint space with $K = 21$ in our case.

Figure 2 depicts the overview of our proposed network architecture. It consists of a 2D pose estimation network (convolutional pose machines - CPM), a CVAE-based statistic network which jointly learns a pose specific embedding from the 3D pose domain and the image feature domain, from which the 3D joint locations are predicted, as well as a novel depth regularizer to provide a weak supervision. Specifically, given a cropped single RGB image containing human hand with certain gesture, we aim to get the 2D heatmaps and the corresponding depth of each joint from the proposed end-to-end network. The 2D joint locations

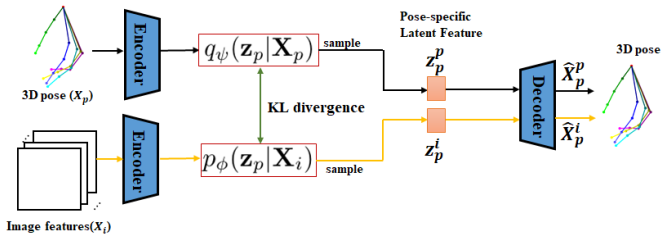


Fig. 3: A pictorial representation of the implemented CVAE-based statistical network, which jointly learns the pose-specific distribution across the paired 3D pose and image feature, from which we sample the latent representation and then decode the predicted 3D hand pose. Note that during training, both 3D labels and image features are utilized as the input of the network, while during testing, we only go through the orange flow to predict the 3D hand pose.

are represented as $\mathbf{\Phi}_{2D} \in \mathbf{\Lambda}_{2D}$, where $\mathbf{\Lambda}_{2D} \in \mathbb{R}^{K\times 2}$ and the depth values are denoted as $\mathbf{\Phi}_z \in \mathbf{\Lambda}_z$, where $\mathbf{\Lambda}_z \in \mathbb{R}^{K\times 1}$. The final output of 3D joint locations are represented in the camera coordinate system, where the first two coordinates are converted from the image plane coordinates using the camera intrinsic matrix, and the third coordinate is the joint depth. We note that although the CVAE-based network is able to decode the 3D hand joint locations and we only choose the z-dimensional result $\mathbf{\Phi}_z$ for our final output, since the 2D projections estimated from the 2D heatmaps slightly outperform those from the CVAE-based network. Moreover, we adopt the implementation of decoding 3D hand joint locations instead of directly regressing the depth of each joint, due to the fact that decoding 3D hand pose in practice brings more robust and accurate estimation of z-dimensional results.

The depth regularizer is the key part to facilitate the proposed weakly supervised training, *i.e.*, relieving the painful joint annotations for real-world dataset by making use of the rough depth maps, which can be easily captured by consumer-grade depth cameras. Additionally, we emphasize that our depth regularizer is only utilized during training. During testing, real images simply go through the dash-dotted line box illustrated in Figure 2 to estimate the 3D hand pose.

The entire network is trained with a Rendered Hand Pose Dataset (RHD) created by [27] and a real-world dataset from Stereo Hand Pose Tracking Benchmark (STB) [41]. For ease of representation, the synthesized dataset and the real-world dataset are denoted as I_{RHD} and I_{STB}, respectively. Note that for weakly-supervised learning, our model is firstly pretrained on I_{RHD} and then adapted to I_{STB} by fusing the training of both datasets. For fully-supervised and semi-supervised learning, the two datasets are used independently in the training and evaluation process.

3.2 2D Pose Estimation Network

For 2D pose estimation, we adopt the encoder-decoder architecture similar to the Convolutional Pose Machines proposed by Wei *et al.* [80], which is fully convolutional with successively refined heatmaps in resolution. The network outputs K low-resolution heatmaps. The intensity on each heat-map indicates the confidence of a joint locating in the 2D position. Here we predict the 2D position of each joint by applying the MMSE (Minimum mean square error given a posterior) estimator, which can be viewed as taking the integration of all locations weighed by their probabilities in the heat map, as proposed in [81]. We initialize the network with weights pretrained by Zimmermann *et al.* [27] and finetune them on I_{RHD}.

To train this module, we employ mean square error (or L2 loss) between the predicted heat map $\hat{\mathbf{\Phi}}_{HM} \in \mathcal{R}^{H\times W}$ and the ground-truth Gaussian heat map $G(\mathbf{\Phi}_{2D}^{gt})$ generated from ground truth 2D labels $\mathbf{\Phi}_{2D}^{gt}$ with standard deviation $\sigma = 1$. The loss function is

$$L_{2D}(\hat{\mathbf{\Phi}}_{HM}, \mathbf{\Phi}_{2D}^{gt}) = \sum_{h}^{H}\sum_{w}^{W}(\hat{\mathbf{\Phi}}_{HM}^{(h,w)} - G(\mathbf{\Phi}_{2D}^{gt})^{(h,w)})^2. \quad (1)$$

3.3 CVAE-based Statistical Network

After obtaining the 2D joint predictions in the form of heatmaps, we aim to infer the 3D hand pose. Most previous work [27], [34], [82] in 3D human pose and hand pose estimation from a single image attempt to lift the set of 2D heatmaps into 3D space directly, while a key issue for this strategy is how to distinguish multiple 3D poses inferred from a single 2D skeleton. Inspired by [33], our method exploits contextual information to reduce the ambiguity of lifting 2D heatmaps to 3D locations. More specifically, we concatenate the intermediate image evidence extracted from the 2D pose estimation network with the predicted 2D heatmaps, and input the concatenated image features to our CVAE-based statistical network. (See supplementary material for more details.)

3.3.1 Conditional Variational Auto-encoder

Given the image features $\mathbf{X}_i$ containing the 2D heatmaps and the intermediate feature representations, we attempt to estimate the corresponding 3D hand pose $\hat{\mathbf{X}}_p$. Mathematically, our goal is to maximize the posterior probability $p(\hat{\mathbf{X}}_p|\mathbf{X}_i)$.

To address this problem, we resort to a CVAE (conditional variational auto-encoder) [79] based framework, whereby a low-dimensional latent representation $\mathbf{z}_p$ is drawn from the prior distribution $p(\mathbf{z}_p|\mathbf{X}_i)$ and the predicted 3D pose $\hat{\mathbf{X}}_p$ is then generated from the distribution $p(\hat{\mathbf{X}}_p|\mathbf{z}_p, \mathbf{X}_i)$. Similar to the original derivation, we start with the posterior formulation that we wish to maximize:

$$p(\hat{\mathbf{X}}_p|\mathbf{X}_i) = \int p_\phi(\hat{\mathbf{X}}_p|\mathbf{X}_i, \mathbf{z}_p)p_\phi(\mathbf{z}_p|\mathbf{X}_i)d\mathbf{z}_p, \quad (2)$$

which can be re-written as follows:

$$\begin{aligned} p(\hat{\mathbf{X}}_p|\mathbf{X}_i) &= \int p_\phi(\hat{\mathbf{X}}_p|\mathbf{X}_i, \mathbf{z}_p)\frac{p_\phi(\mathbf{z}_p|\mathbf{X}_i)}{q_\psi(\mathbf{z}_p|\mathbf{X}_p)}q_\psi(\mathbf{z}_p|\mathbf{X}_p)d\mathbf{z}_p \\ &= E_{\mathbf{z}_p\sim q_\psi(\mathbf{z}_p|\mathbf{X}_p)}\{p_\phi(\hat{\mathbf{X}}_p|\mathbf{X}_i, \mathbf{z}_p)\frac{p_\phi(\mathbf{z}_p|\mathbf{X}_i)}{q_\psi(\mathbf{z}_p|\mathbf{X}_p)}\}. \end{aligned} \quad (3)$$

After taking logs and applying Jensen's inequality, we obtain the variational lower bound of the conditional log-likelihood of the observation:

$$\begin{aligned} log(p(\hat{\mathbf{X}}_p|\mathbf{X}_i)) &\geq E_{\mathbf{z}_p\sim q_\psi(\mathbf{z}_p|\mathbf{X}_p)}[log(p_\phi(\hat{\mathbf{X}}_p|\mathbf{X}_i,\mathbf{z}_p)) \\ &+ log(\frac{p_\phi(\mathbf{z}_p|\mathbf{X}_i)}{q_\psi(\mathbf{z}_p|\mathbf{X}_p)})] \\ &= E_{\mathbf{z}_p\sim q_\psi(\mathbf{z}_p|\mathbf{X}_p)} log(p_\phi(\hat{\mathbf{X}}_p|\mathbf{X}_i,\mathbf{z}_p)) \\ &- \mathbf{KL}(q_\psi(\mathbf{z}_p|\mathbf{X}_p)||p_\phi(\mathbf{z}_p|\mathbf{X}_i)). \end{aligned} \tag{4}$$

Here $\mathbf{KL}$ is the Kullback-Leibler divergence, $\mathbf{z}_p$ the pose-specific latent embedding, $\mathbf{X}_p$ the ground truth 3D label, $q_\psi(\mathbf{z}_p|\mathbf{X}_p)$ the posterior sampling function, $p_\phi(\mathbf{z}_p|\mathbf{X}_i)$ the conditional prior, and $p_\phi(\hat{\mathbf{X}}_p|\mathbf{z}_p,\mathbf{X}_i)$ is the likelihood. Note that the conditional probability distributions ψ,ϕ can be parameterized by deep neural networks.

For our purposes, we assume that the low dimensional subspace serves as a controller for the dependency of the high-dimensional 3D hand configurations, which indicates that the extracted pose-specific latent feature z_p encoded from the 3D annotations $\mathbf{X}_p$ is sufficient to infer the 3D hand pose $\hat{\mathbf{X}}_p$. Therefore, $p_\phi(\hat{\mathbf{X}}_p|\mathbf{z}_p,\mathbf{X}_i)$ can be replaced by $p_\phi(\hat{\mathbf{X}}_p|\mathbf{z}_p)$ and Eq. (4) can be updated as:

$$\begin{aligned} log(p(\hat{\mathbf{X}}_p|\mathbf{X}_i)) \geq &- \mathbf{KL}(q_\psi(\mathbf{z}_p|\mathbf{X}_p)||p_\phi(\mathbf{z}_p|\mathbf{X}_i)) \\ &+ E_{q_\psi(\mathbf{z}_p|\mathbf{X}_p)}[log(p_\phi(\hat{\mathbf{X}}_p|\mathbf{z}_p))]. \end{aligned} \tag{5}$$

However, the second term in Eq. (5) only takes the sampling from $q_\psi(\mathbf{z}_p|\mathbf{X}_p)$ during training, which is encoded by the 3D annotations. This may not be optimal to make a prediction during testing, as we wish to reconstruct 3D hand pose typically from the extracted image features $\mathbf{X}_i$. In order to mitigate the gap between training and testing processes, inspired from [79], we modify Eq. (5) by setting the reconstruction with sampling from both $q_\psi(\mathbf{z}_p|\mathbf{X}_p)$ and $p_\phi(\mathbf{z}_p|\mathbf{X}_i)$, and we have the following notation:

$$\begin{aligned} log(p(\hat{\mathbf{X}}_p|\mathbf{X}_i)) \geq &\mu\{-\mathbf{KL}(q_\psi(\mathbf{z}_p|\mathbf{X}_p)||p_\phi(\mathbf{z}_p|\mathbf{X}_i))\} \\ &+ \mu E_{q_\psi(\mathbf{z}_p|\mathbf{X}_p)}[log(p_\phi(\hat{\mathbf{X}}_p|\mathbf{z}_p))] \\ &+ (1-\mu) E_{p_\phi(\mathbf{z}_p|\mathbf{X}_i)}[log(p_\phi(\hat{\mathbf{X}}_p|\mathbf{z}_p))]. \end{aligned} \tag{6}$$

where $0 \leq \mu \leq 1$, which balances the two objectives during training stage. Here we set $\mu = 0.5$ in our experiment.

3.3.2 Network Architecture and Loss function

A pictorial representation of the implemented CVAE-based statistical network is provided in Figure 3, which consists of two paths. For the upper path, the ground truth 3D label $\mathbf{X}_p$ is leveraged to encode the distribution $q_\psi(\mathbf{z}_p|\mathbf{X}_p)$ during training, while for the lower path, which is also the test path, the latent distribution $p_\phi(\mathbf{z}_p|\mathbf{X}_i)$ is inferred based on the extracted image features $\mathbf{X}_i$. Both $q_\psi(\mathbf{z}_p|\mathbf{X}_p)$ and $p_\phi(\mathbf{z}_p|\mathbf{X}_i)$ can be sampled to reconstruct the 3D hand pose.

To train the CVAE-based framework, we propose the loss function with the combination of terms in Eq. (6) to maximize the variational lower bound and minimize the total loss:

$$L_{cvae} = L_{KL} + L_{pose}, \tag{7}$$

where L_{KL} represents the KL divergence between pairs of latent distributions from image feature domain and 3D pose domain, and L_{pose} denotes the reconstruction loss of the predicted 3D hand pose. Next, we will provide more details on each term, with intuitive interpretations separately.

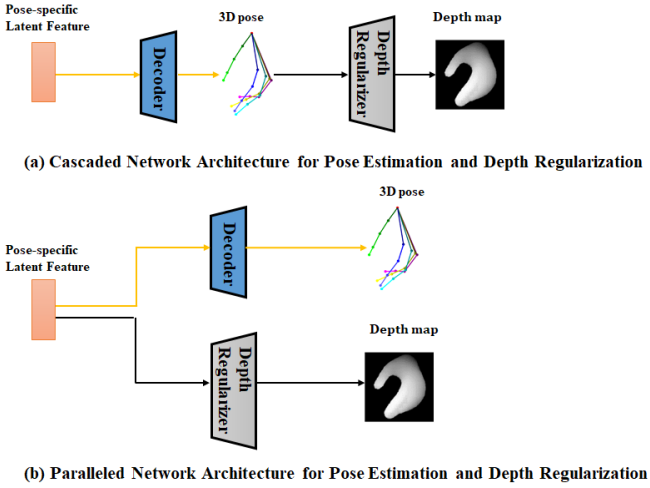


Fig. 4: Two architectures for using the depth regularizer, either (a) cascaded or (b) paralleled with the 3D hand pose decoder. Such a depth regularizer takes the easily-captured reference depth map to provide weak supervision for unlabeled data. Note that the depth regularizer is only leveraged during training. During testing, image features go through the orange flow to predict the 3D hand pose, as shown in both (a) and (b).

The distribution regularization L_{KL} aims to regularizes consistency between pairs of the pose-specific distributions $q_\psi(\mathbf{z}_p|\mathbf{X}_p)$ and $p_\phi(\mathbf{z}_p|\mathbf{X}_i)$, in terms of KL divergences, which ensures the latent distribution encoded from image feature domain close to that from the 3D hand pose domain. Practically, to simplify the computation, we define both $q_\psi(\mathbf{z}_p|\mathbf{X}_p)$ and $p_\phi(\mathbf{z}_p|\mathbf{X}_i)$ as Gaussian distribution.

$$L_{KL} = -\mu\{\mathbf{KL}(q_\psi(\mathbf{z}_p|\mathbf{X}_p)||p_\phi(\mathbf{z}_p|\mathbf{X}_i))\}. \tag{8}$$

L_{pose} can be interpreted as the likelihood terms in Eq. (6), encouraging accurate 3D hand pose reconstructions from 3D pose and RGB image domains. Particularly, for a collection of training datasets, the likelihood terms can be approximated by drawing samples $\mathbf{z}_p^{(l)}(l = 1,2,...L)$, $\mathbf{z}_p^{(m)}(m = 1,2,...M)$ from the pose specific distributions $q_\psi(\mathbf{z}_p|\mathbf{X}_p)$ and $p_\phi(\mathbf{z}_p|\mathbf{X}_i)$, respectively:

$$E_{q_\psi(\mathbf{z}_p|\mathbf{X}_p)}[log(p_\phi(\hat{\mathbf{X}}_p|\mathbf{z}_p))] = \frac{1}{L}\sum_{l=1}^{L} log(p_\phi(\hat{\mathbf{X}}_p|\mathbf{z}_p^{(l)})) \tag{9}$$

$$E_{p_\phi(\mathbf{z}_p|\mathbf{X}_i)}[log(p_\phi(\hat{\mathbf{X}}_p|\mathbf{z}_p))] = \frac{1}{M}\sum_{m=1}^{M} log(p_\phi(\hat{\mathbf{X}}_p|\mathbf{z}_p^{(m)})). \tag{10}$$

Then, L_{pose} is defined as:

$$L_{pose} = \mu smooth_{L1}(\hat{\mathbf{X}}_p^p - \mathbf{X}_p) + (1-\mu) smooth_{L1}(\hat{\mathbf{X}}_p^i - \mathbf{X}_p), \tag{11}$$

where $smooth_{L1}$ denotes the smooth L1 loss introduced in [83], $\hat{\mathbf{X}}_p^p = Decoder(\mathbf{z}_p^p)$, $\hat{\mathbf{X}}_p^i = Decoder(\mathbf{z}_p^i)$, $\mathbf{z}_p^p$ and $\mathbf{z}_p^i$ are the latent features sampled from $q_\psi(\mathbf{z}_p|\mathbf{X}_p)$ and $p_\phi(\mathbf{z}_p|\mathbf{X}_i)$, as shown in Figure 3, and $\mathbf{X}_p$ is the ground truth 3D hand pose.

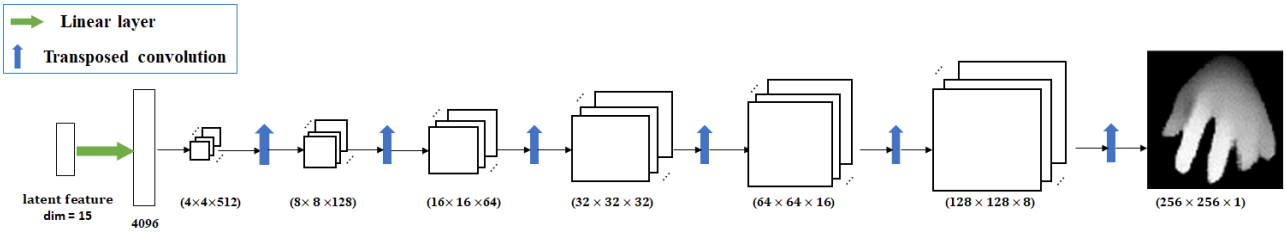


Fig. 5: Network architecture of our proposed depth regularizer. Given pose-specific latent feature as the input, the depth regularizer is able to render the corresponding depth map by gradually enlarging the intermediate feature maps and finally combining them into a single depth image.

3.4 Depth Regularizer

For fully-supervised learning with 3D annotations, the above mentioned CVAE-based statistical network is sufficient for both training and testing process, since we use 3D labels to guide the shared latent space and 3D pose predictions. However, for weakly-supervised learning with unlabeled real-world images, no penalty can be enforced because of the absence of 3D annotations.

To address this issue, given the fact that the easily-captured depth images can be applied as an implicit constraint of physical structures, we introduce a novel depth regularizer leveraging low-cost reference depth maps to provide a weak supervision for 3D hand pose. More specifically, inspired by [38], [84] that iteratively synthesize depth images from 3D pose, we deploy a deep neural network to render depth images directly from the corresponding 3D hand pose, which is referred as cascaded depth regularizer, as shown in Figure 4 (a). Since 3D hand pose can be decoded from a pose-specific latent feature, we assume that when tuning the pose-specific feature through the depth regularizer, the decoded 3D hand pose can be adapted as well. Moreover, we present another strategy for the network layout, which facilitates the depth regularizer during training by constraining the depth maps generated from the low-dimensional pose-specific features, *i.e.*, the proposed depth regularizer is in parallel with the 3D hand pose decoder, as illustrated in Figure 4 (b). In Sec. 4.3.4, we experimentally show that our proposed paralleled architecture outperforms the cascaded one, which suggests the benefits of constraining the low-dimensional latent feature.

To train the depth regularizer, we adopt L1 norm which is more robust for image generation task compared with L2 norm, to minimize the difference between the ground truth depth map $\mathbf{X}_d$ and the depth images $\hat{\mathbf{X}}_d^i$, $\hat{\mathbf{X}}_d^p$ generated from the latent feature $\mathbf{z}_p^i$, $\mathbf{z}_p^p$, respectively:

$$L_{dep}(\hat{\mathbf{X}}_d^i, \hat{\mathbf{X}}_d^p, \mathbf{X}_d) = ||\hat{\mathbf{X}}_d^p - \mathbf{X}_d||_1 + ||\hat{\mathbf{X}}_d^i - \mathbf{X}_d||_1. \quad (12)$$

3.5 Training

For weakly-supervised learning, similar to [33] and [85], we adopt fused training where each mini-batch contains both the synthetic and the real-world training examples (half-half), shuffled randomly during the training process.

To train the whole network, for well-labeled synthetic data, we obtain the overall loss function by combining terms in Eq. (1), (7) and (12):

$$L_s = \lambda_{2D} L_{2D} + \lambda_{cvae} L_{cvae} + \lambda_{dep} L_{dep}. \quad (13)$$

where L_{2D} is responsible for the prediction of 2D heatmap, L_{cvae} arises from the CVAE-based framework, and L_{dep} is the generation loss from our proposed depth regularizer. In contrast, for real-world data without 3D labels, we only utilize the depth regularizer to provide weak supervision. Therefore, the overall loss function is simplified as:

$$L_r = \lambda_{dep} L_{dep}. \quad (14)$$

4 Experiments

4.1 Implementation Details

In this section, we describe the details of the network architecture and implementation details.

We apply simple network in our proposed CVAE-based framework and depth regularizer. More precisely, the encoder for the image features contains two convolutional layers with the kernel size 3, stride 2 and padding 1, followed by a fully connected layer. Moreover, similar to [29], we design the same architecture for the encoder/decoder of 3D hand pose, containing four fully connected layers with 512 hidden units. As for the layout of our proposed depth regularizer, inspired by [38], [86], we deploy a network which passes through a fully-connected layer connected with six convolutional layers, as shown in Figure 5. Each convolutional layer contains a transposed convolution followed by a ReLU, after which the feature map is expanded along both image dimensions. In the first five convolutional layers, batch normalization [87] is introduced before ReLU in order to reduce the dependency on the initialization. After that, the final layer combines all feature maps to generate the corresponding depth map.

Our method is implemented within PyTorch platform. Adam optimizer [88] is used for training. In our experiments, we adopt a three-stage training process, which is more effective in practice compared with direct end-to-end training. In particular, *Stage 1* fine-tunes the 2D pose estimation network with weights from Zimmermann *et al.* [27], which are adapted from the Convolutional Pose Machines [80]. *Stage 2* trains the CVAE-based statistical framework and depth regularizer from scratch with synthetic data. *Stage 3* fine-tunes the whole network with all the training data, which is an end-to-end training. During optimization, the weights of different losses are set to $\lambda_{2D} = 1$, $\mu = 0.5$, $\lambda_{cvae} = 0.1$ and $\lambda_{dep} = 0.1$, with a batch size of 8 and a regularization strength of 5×10^{-4}. All experiments are conducted on one GeForce GTX 1080 GPU with CUDA 8.0.

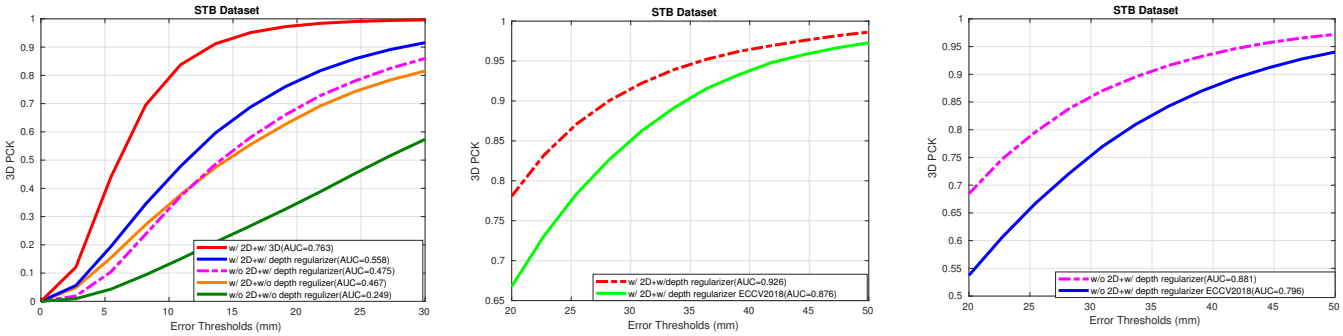


Fig. 6: Left: Comparisons of 3D PCK results of different baselines with our weakly-supervised method on STB [41]. Middle/Right: Comparison of the weakly-supervised method with our conference paper [40] on STB [41] dataset, in scenarios with both depth regularization and 2D supervision (middle) or with only depth regularization (right).

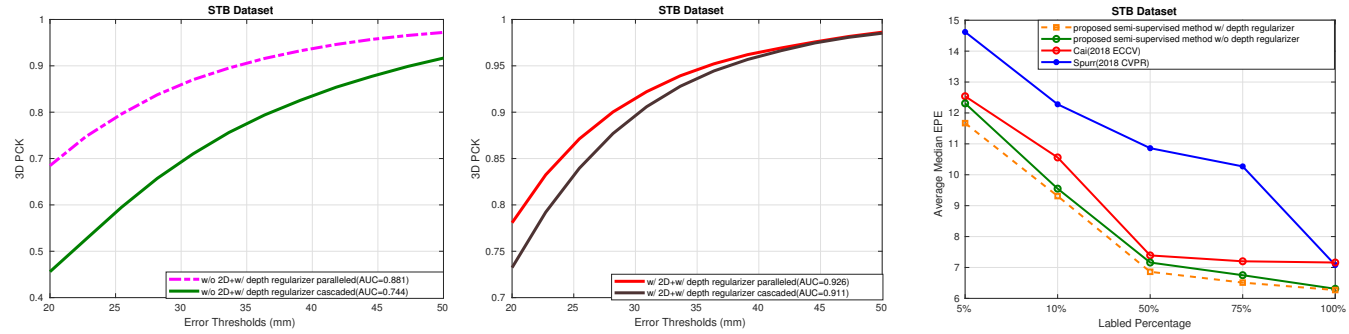


Fig. 7: Left/Middle: Comparisons of the 3D PCK results of different architectures of the depth regularizer for weakly-supervised method. Left: Weakly-supervised flow with only depth regularization. Middle:Weakly-supervised flow with both depth regularization and 2D supervision. Right: Median EPE comparisons with the state-of-the-art methods on STB [41] dataset in supervised and semi-supervised manner as a function of the percentage of labeled data.

4.2 Datasets and Metrics

We evaluate our method mainly on four publicly available datasets: Rendered Hand Pose Dataset (RHD) [27], real-world dataset from Stereo Hand Pose Tracking Benchmark (STB) [41], Egodexter [43] and Dexter+Object [42] datasets, where Egodexter and Dexter+Object are simply used for validation. Note that here we infer a scale-invariant and translation-invariant representation of 3D hand pose, by subtracting from each hand joint the location of root keypoint and then normalizing it by the distance between a certain pair of keypoints, as done in [27], [28].

RHD is a synthetic dataset of rendered hand images with a resolution of 320×320, which is built upon 20 different characters performing 39 actions and is composed of 41,258 images for training and 2,728 images for testing. All samples are annotated with 2D and 3D keypoint locations. For each RGB image, the corresponding depth image is also provided. This dataset is considerably challenging due to the large variations in viewpoints and hand shapes, as well as the large visual diversity induced by random noise and different illuminations. With all the labels provided, we train the entire proposed network, including the 2D pose estimation network, the CVAE-based statistical network and the depth regularizer.

STB is a real world dataset containing two subsets with an image resolution of 640×480: the stereo subset STB-BB captured from a Point Grey Bumblebee2 stereo camera and the color-depth subset STB-SK captured from an active depth camera. Note that the two types of images are captured simultaneously with the same resolution, identical camera pose, and similar viewpoints. Both STB-BB and STB-SK provide 2D and 3D annotations of 21 keypoints. For weakly-supervised experiments, we use color-depth pairs in STB-SK dataset, as well as root depth (*i.e.*, palm in the experiments) and hand scale (the distance between a certain pair of keypoints). For fully-supervised experiments, both color-depth pairs (STB-BB) and stereo pairs (STB-SK) with 2D and 3D annotations are utilized to train the whole network. Note that all experiments conducted on STB dataset follow the same training and evaluation protocol used in [27], [28], which train on 10 sequences and test on the other two.

Dexter+Object [42], in total, provides 6 test video sequences with 3145 frames. All sequences are recorded using a static camera with a single person interacting with an object, where 2D and 3D pose annotations of visible fingertips are provided for the dataset. Similarly, EgoDexter [43] contains 4 testing video sequences with 3190 frames, which are recorded from egocentric viewpoints with cluttered backgrounds and complex hand-object interactions. 2D and 3D pose annotations of visible fingertips are provided for most frames of this dataset. We note that the two validation

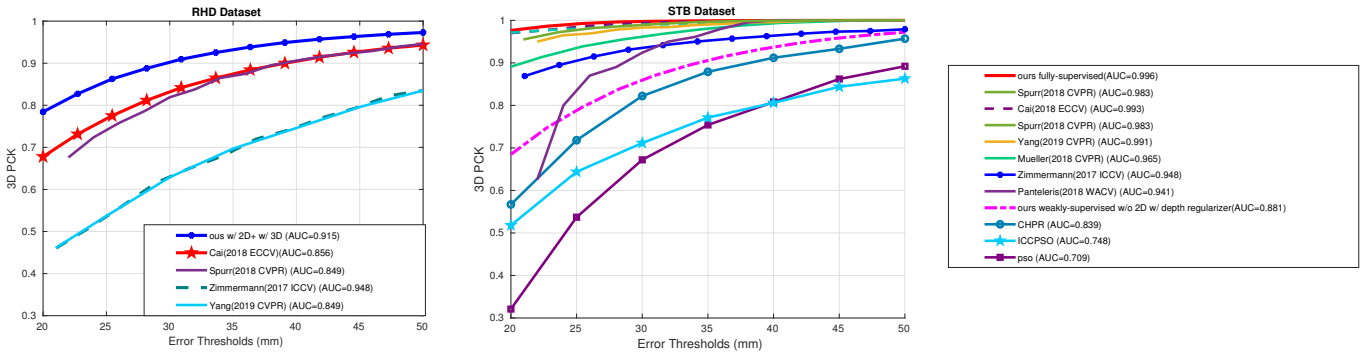


Fig. 8: Comparisons with the state-of-the-art fully-supervised methods on RHD [27] and STB [41]. Left: 3D PCK on RHD dataset. Right: 3D PCK on STB dataset.

datasets contain rich examples of hand-object interactions and egocentric viewpoints, which are hard to find in our training datasets (RHD and STB). To mitigate the gap between training and validation scenarios and improve the robustness of our proposed weakly-supervised model, an extra real-world dataset (CMU Multiview Bootstrapping [89], referred as CMUMB) with only 2D supervision is added to the training process when measuring the estimation performance of the two datasets.

We evaluate the 3D hand pose estimation performance with three metrics. The first metric is the area under the curve (AUC) on the percentage of correct keypoints (PCK) score, which is a popular criterion to evaluate the pose estimation accuracy with different thresholds, as proposed in [27], [28]. The second and third metrics are the mean and median end-point-error (EPE) over all testing frames, as utilized in [29]. Following the same condition used in [27], [28] for STB and RHD datasets, we assume that the global hand scale and the root depth are known in the experimental evaluations so that we can report the quantitative results based on 3D hand joint location in the global coordinate frame, which are computed from the output root-relative articulations. For Dexter Object and Egodexter datasets, we follow [60] to calculate the absolute 3D pose with global scale using the predicted normalized root-relative 3D pose and the intrinsic camera parameters.

4.3 Self-Comparison

4.3.1 Weak Supervision

We first evaluate the impact of weak label constraints on STB dataset compared with fully-supervised methods with complete 2D and 3D annotations. Specifically, we compare our proposed weakly-supervised approach (**w/o 2D + w/ depth regularizer**) with four baselines: a) **w/o 2D + w/o depth regularizer**: directly using the network model pretrained on RHD dataset; b) **w/ 2D + w/o depth regularizer**: tuning the pretrained network with 2D labels on STB dataset; c) **w/ 2D + w/ depth regularizer**: tuning the pretrained network with both 2D labels and depth regularizer on STB dataset and d) **w/ 2D + w/ 3D**: fully-supervised method with complete 2D and 3D annotations.

As illustrated in the left part of Figure 6, the fully-supervised method (baseline-d) achieves the best performance while directly transferring the model trained on synthetic data with no adaptation (baseline-a) yields the worst estimation results. This is not surprising, since the fully-supervised method provides the most effective constraint in the 3D hand pose estimation task and real-world images have considerable domain shift from synthetic ones. Note that these two baselines serve as upper bound and lower bound for our weakly-supervised method, respectively. Compared with baseline-a, by fine-tuning the pretrained model with the 2D labels of the real images (baseline-b), the AUC value significantly increases from 0.249 to 0.467. Moreover, as shown in Figure 6, adding our proposed depth regularizer without 2D supervision (our proposed weakly-supervised approach) achieves better performance than baseline-b, increasing AUC to 0.475, which suggests that the depth regularizer provides even stronger supervision than 2D labels. Furthermore, baseline-c with both 2D labels and depth regularizer further improves the AUC value to 0.558, demonstrating that the network can benefit from the complementarity between depth regularizer and 2D supervision.

We also compare the weakly-supervised method with our conference paper [40]. As shown in Figure 6 (middle and right), our proposed model surpasses the accuracy of [40], which improves the AUC value from 0.796 to 0.881 for weak-supervised method with only depth regularizer. Moreover, in scenarios with both 2D labels and depth regularization, our proposed method outperforms [40] as well, increasing the AUC value from 0.876 to 0.926. It is worth noting that our proposed weakly-supervised method with only depth regularization (AUC value: 0.881) even performs comparably with the original method [40] with both 2D labels and depth regularization (AUC value: 0.876), which further indicates the advanced generalization ability of the proposed pipeline.

4.3.2 Semi-supervision

To further assess the effectiveness of our proposed depth regularizer, we also explore the semi-supervised situation, where both labeled data with 3D annotations and unlabeled data with depth maps are available during training process. This is a common scenario, as unlabeled RGB images with depth maps can be easily captured by commodity RGB-D cameras.

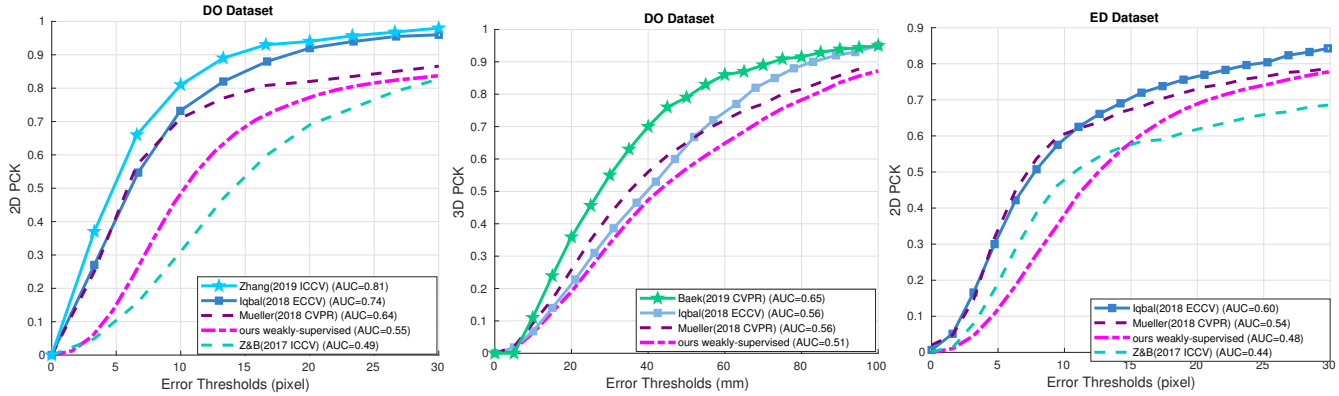


Fig. 9: Comparisons with the state-of-the-art methods on Dexter Object [42] and Egodexter [43]. Left: 2D PCK on Dexter Object dataset. Middle: 3D PCK on Dexter Object dataset. Right: 2D PCK on Egodexter dataset.

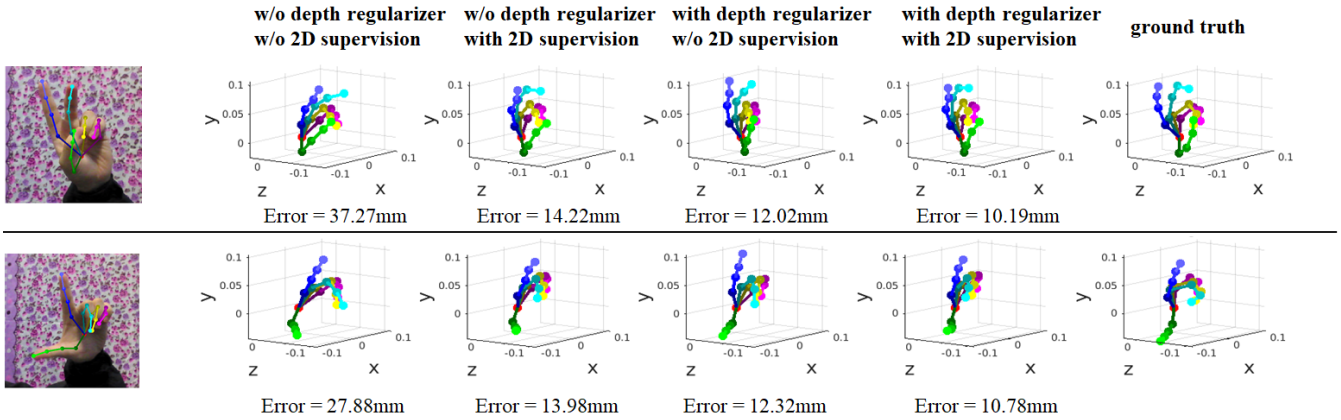


Fig. 10: Visual results of our proposed weakly-supervised approach trained with only depth regularization (column 1, 4) and with both 2D labels and depth regularization (column 5), as well as other baselines (non-finetuned approach in column 2, and weakly-supervised method with only 2D supervision in column 3), compared with the ground truth 3D hand pose (column 6). Note that columns 2-6 are shown at a novel viewpoint for easy comparison.

TABLE 1: Impact of our proposed depth regularizer on 3D hand pose estimation with different percentages of labeled data. The Mean EPE (mm) on STB dataset are listed in this table.

	0%	5%	10%	50%	75%	100%
w/o depth reg.	30.18	13.95	11.61	8.11	7.58	7.12
w/ depth reg.	**17.61**	**13.33**	**11.36**	**7.80**	**7.30**	**7.10**

TABLE 2: Effectiveness of our proposed CVAE-based framework compared with the original regression network in [40] on 3D hand pose estimation with different percentages of labeled data. Note that for fair comparison, depth regularizer is not utilized in both strategies. The Mean EPE (mm) on STB dataset are listed in this table.

	5%	10%	50%	75%	100%
[40] regression network	14.25	12.22	8.37	8.13	7.73
proposed CVAE framework	**13.95**	**11.61**	**8.11**	**7.58**	**7.12**

Table 1 compares the results of the experiments with different percentages of 3D joint labels in STB dataset. We see that with more unlabeled data used in the dataset, our proposed depth regularizer leads to larger gain in 3D hand pose estimation, which indicates the regularizing effect of the additional depth maps. Specifically, we note that for fully-supervised method (100% in Table 1), the individual performance difference brought by our proposed depth regularizer is minor, while on weakly-supervised setting (0% in Tabel 1), the margin is much more significant, up to 12.57mm. This can be expected, as we already leverage the strong supervision of 3D annotations in fully-supervised method, which represents much more accurate 3D hand pose information than depth maps. In contrast, for weakly-supervised task, the complementary depth information becomes significant for closing the large domain shift between synthetic and real-world dataset, so as to considerably improve the performance in a relatively large margin.

4.3.3 CVAE-based framework Versus Simple Regression Network

We also compare the influence of our proposed CVAE-based statistical framework with the simple regression network presented in our conference paper [40], which is illustrated in Table 2. For fair comparison, we do not use depth regularizer in our experiment. Therefore, Mean EPE of the weakly-supervised method (0% of labeled data) is not reported in

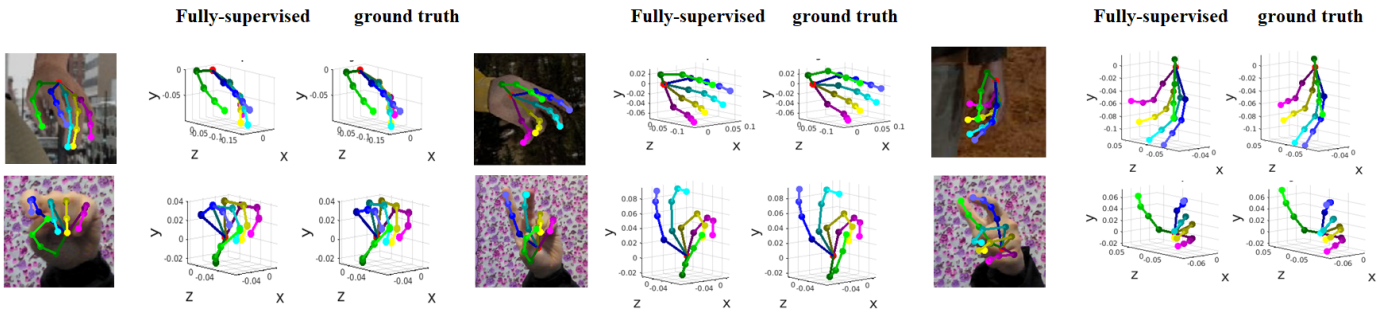


Fig. 11: Visual results of our fully-supervised method on RHD and STB datasets. Row 1-2: RHD dataset. Row 3-4: STB dataset. Note that skeletons are shown at a novel viewpoint for easy comparison.

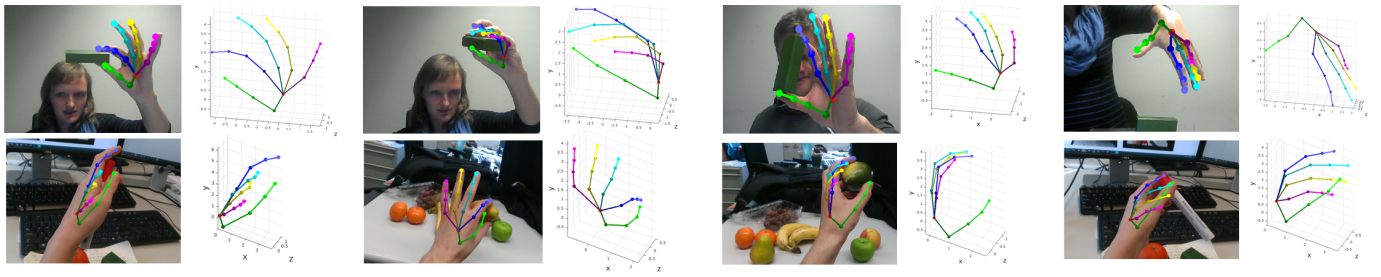


Fig. 12: Qualitative Results for Dexter Object [42](row 1-2) and Egodexter [43] (row 3-4) datasets for weakly-supervised approach. An additional CMUVB dataset [89] with only 2D supervision is added during training to compensate the absence of hand-object and egocentric samples in original training datasets (STB and RHD). Note that skeletons are shown at a novel viewpoint for easy comparison.

the table. It is worth noting that, when using CVAE-based framework, our method outperforms the original pipeline in both semi-supervised and fully-supervised method (100% of labeled data), as shown in Table 2. In Section 4.5, we show extensive qualitative comparisons for the weakly-supervised flow with only 2D supervision, which demonstrates the advantages of the CVAE-based framework in producing plausible physical structures of 3D hand pose.

4.3.4 *Cascaded Depth Regularizer Versus Paralleled one*

In this section, we investigate the prediction accuracy of our proposed depth regularizer operated on pose-specific latent representation (paralleled network architecture) or directly on 3D pose estimation (cascaded network architecture), as described in Sec. 3.4. For fair comparison, as shown in Figure 4, we experiment with similar network architecture, where the only difference is the layout of the depth regularizer, either cascaded or paralleled with the 3D hand pose decoder. Figure 7 (left and middle) reports the comparative results for weakly-supervised method on STB dataset, which shows that the paralleled depth regularizer is superior to the cascaded one, indicating that the network can benefit more from directly constraining the low-dimensional pose-specific latent feature, compared with constraining the 3D hand pose.

4.4 Comparison with State-of-the-art Methods

4.4.1 *Fully-supervised and Weakly-supervised Methods*

Figure 8 shows the comparisons with state-of-the-art methods, including PSO [49], ICPPSO [90], Panteleris *et al.* [61], Zimmermann *et al.* [27], Mueller *et al.* [28], Spurr *et al.* [29], Iqbal *et al.* [60], Yang *et al.* [78] and our conference paper [40] on both RHD and STB datasets. Note that here we report the performance of our fully-supervised method without depth regularizer.

As shown in Figure 8 (left), on RHD dataset, our fully-supervised method is superior to the state-of-the-art methods. The AUC value of our proposed method is 0.24, 0.066, 0.059 and 0.059 higher than those of the methods in [27], [29], [40], [78], respectively.

On STB dataset, our fully-supervised method achieves the best results compared with all existing methods, improving the AUC value to 0.996 in joint error range between 20mm and 50mm. Note that our weakly-supervised method with only depth regularizer also outperforms some of the existing fully-supervised methods, which demonstrates the potential values for the weakly-supervised exploration when complete 3D annotations are difficult to obtain in real-world dataset. We would like to point out that the gap between the weakly-supervised and fully-supervised approaches is partially from the skeleton differences between synthetic and real data, as different annotation schemes are adopted by the synthetic RHD [27] and real-world STB [41] datasets. It is also noted that the AUC values of our proposed methods in Figure 8 (right) are slightly different from their counterparts in Section 4.3.1. This is because here we test on the stereo pair subset STB-BB rather than the color-depth subset STB-SK.

To further assess the generalization quality of our proposed weakly-supervised approach, we additionally provide quantitative analysis on Dexter Object [42] and

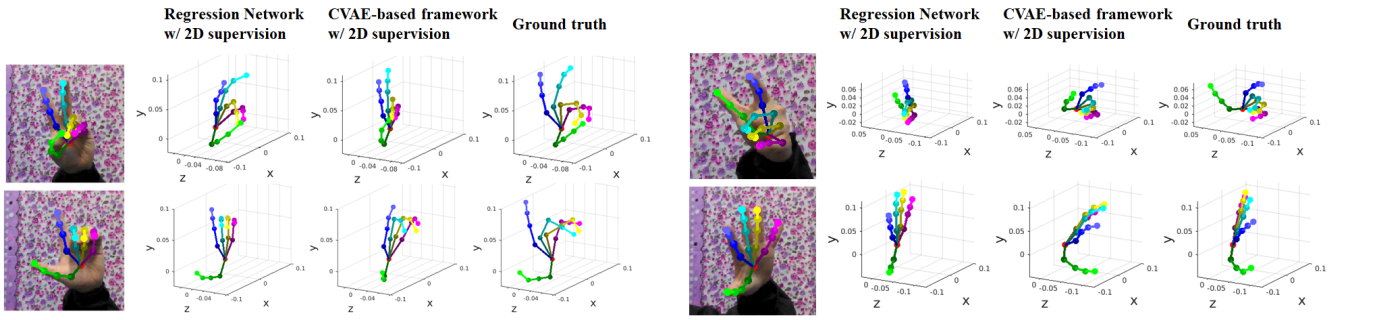


Fig. 13: Visual results of our proposed CVAE-based statistical framework (column 1, 3) and the original regression framework (column 2) presented in our conference paper [40]. Here both methods are trained on STB [41] dataset with only 2D supervision and without depth regularization. Skeletons in column 2-4 are shown at a novel viewpoint for easy comparison.

Egodexter [43] datasets, as shown in Fig 9. Note that different from other state-of-the-art approaches [27], [28], [60], [66], [70] mentioned in Fig 9, our proposed weakly-supervised method didn't take real-world 3D labels of other datasets as supervision during training. We can see that our weakly-supervised approach is comparable with other fully-supervised approaches with real-world 3D annotations, which indicates the generalization ability of our approach to real-world scenarios.

4.4.2 *Semi-supervised Methods*

We also perform a quantitative analysis of our semi-supervised method on STB dataset with the state-of-the-art semi-supervised methods [29], [40] for RGB-based 3D hand pose estimation. As presented in Figure 7 (right), our proposed semi-supervised approach without depth regularizer outperforms [29] by a relatively large margin. Additionally, the proposed CVAE-based model consistently surpasses our conference paper [40], which validates that our proposed CVAE-based statistical framework can better capture the 3D pose information and provide more accurate estimation, compared with the original regression network. Furthermore, Figure 7 (right) also shows that adding depth regularizer can further improve the estimation accuracy of the semi-supervised method, as discussed in Section 4.3.2.

4.5 Qualitative Results

Figure 10 shows some visual results of our proposed weakly-supervised approach and baselines. For a better comparison, we show the 3D skeleton reconstructions at a novel view and the skeleton reconstructions of our method at the original view are overlaid with the input images. It can be seen that, directly applying the model pretrained on synthetic dataset to real dataset (column 2) fails to capture the global orientation of 3D pose, while finetuning with 2D labels improve this situation to some extent, as shown in column 3. It is worth noting that after simply imposing the depth regularizer without 2D supervision (column 1, 4), both the 2D estimation results and global orientations considerably perform better, compared with the non-finetuned method in column 2. Finally, with both 2D constraint and depth regularizer, the weakly-supervised approach on real-world dataset yields the best estimation accuracy, which is consistent with our aforementioned quantitative analysis.

Figure 11 shows some visual results of our fully-supervised methods on RHD and STB datasets. We exhibit samples captured from various viewpoints with serious self-occlusions. It can be seen that our fully-supervised approach is robust to various hand orientations and complicated pose articulations.

To evaluate the generalization ability of our proposed weakly-supervised method, we also provide qualitative examples on another two challenging datasets (Egodexter [43] and Dexter Object [42]), as shown in Figure 12. As can be seen in Figure 12, our method is able to provide valid pose estimation for challenging scenarios in daily life.

We also conduct qualitative comparison between our proposed CVAE-based statistical framework and the regression network in our original paper [40]. For a fair comparison, we train the network on STB dataset with only 2D supervision. Note that no depth regularizer is introduced in this setting. Qualitative results are provided in Figure 13. As can be seen, although our proposed method did not accommodate well with the global orientation due to the absence of depth supervision, it still performs better in exploiting plausible and valid 3D pose structures than the original simple regression network.

5 Conclusion

Providing full 3D annotations for a large real-world hand dataset is a major bottleneck for learning-based 3D hand pose. To address this problem, we have presented a weakly supervised solution, which can leverage unlabeled real-world dataset by adapting fully-annotated synthetic dataset with the aid of low-cost depth images, which, to our knowledge, is the first exploration of leveraging depth maps to compensate the absence of entire 3D annotations for 3D hand pose estimation from RGB images. To be specific, we have introduced a novel and effective end-to-end architecture consisting of a 2D estimation network, a CVAE-based statistical network, and a novel depth regularizer. Quantitative and qualitative experimental results have demonstrated that our weakly-supervised method compares favorably with the existing works, and our semi-supervised and fully-supervised approaches surpass the state-of-the-art methods.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This research is also supported in part by Singapore MoE Tier-2 Grant (MOE2016-T2-2-065) and start-up funds from Monash University and University at Buffalo.

REFERENCES

[1] J. M. Rehg and T. Kanade, "Digiteyes: Vision-based hand tracking for human-computer interaction," in *Pro. IEEE Workshop Motion of Non-rigid and Articulated Objects*, 1994, pp. 16–22.

[2] Y. Wu and T. S. Huang, "Capturing articulated human hand motion: A divide-and-conquer approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1, 1999, pp. 606–611.

[3] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Trans. Graph.*, vol. 33, no. 5, p. 169, 2014.

[4] D. Tang, H. Jin Chang, A. Tejani, and T.-K. Kim, "Latent regression forest: Structured estimation of 3d articulated hand posture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3786–3793.

[5] H. Liang, J. Yuan, and D. Thalmann, "Resolving ambiguous hand pose predictions by exploiting part correlations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 7, pp. 1125–1139, 2015.

[6] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei *et al.*, "Accurate, robust, and flexible real-time hand tracking," in *Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst.*,. ACM, 2015, pp. 3633–3642.

[7] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff, "3d hand pose reconstruction using specialized mappings," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1, 2001, pp. 378–385.

[8] Y. Wu and T. S. Huang, "View-independent recognition of hand postures."

[9] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, 2018.

[10] J. Weng, C. Weng, J. Yuan, and Z. Liu, "Discriminative spatio-temporal pattern discovery for 3d action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, 2018.

[11] J. Weng, C. Weng, and J. Yuan, "Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4171–4180.

[12] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Hand pose estimation and hand shape classification using multi-layered randomized decision forests," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 852–863.

[13] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

[14] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "3d convolutional neural networks for efficient and robust hand pose estimation from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2017, p. 5.

[15] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman *et al.*, "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2821–2840, 2013.

[16] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1385–1392.

[17] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3593–3601.

[18] Y. Zhou, J. Lu, K. Du, X. Lin, Y. Sun, and X. Ma, "Hbe: Hand branch ensemble network for real-time 3d hand pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 501–516.

[19] X. Wu, D. Finnegan, E. O'Neill, and Y.-L. Yang, "Handmap: Robust hand pose estimation via intermediate dense guidance map supervision," in *Proc. Eur. Conf. Comput. Vis.* Springer, Cham, 2018, pp. 246–262.

[20] Q. Ye and T.-K. Kim, "Occlusion-aware hand pose estimation using hierarchical mixture density network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–817.

[21] J. Malik, A. Elhayek, F. Nunnari, K. Varanasi, K. Tamaddon, A. Heloir, and D. Stricker, "Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth," in *Proc. Int. Conf. 3D Vis.*, 2018, pp. 110–119.

[22] P. Li, H. Ling, X. Li, and C. Liao, "3d hand pose estimation using randomized decision forest with segmentation index points," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 819–827.

[23] M. Oberweger, G. Riegler, P. Wohlhart, and V. Lepetit, "Efficiently creating 3d training data for fine hand pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4957–4965.

[24] M. Rad, M. Oberweger, and V. Lepetit, "Feature mapping for learning fast and accurate 3d pose inference from synthetic images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4663–4672.

[25] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 824–832.

[26] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim, "Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4866–4874.

[27] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017.

[28] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2018. [Online]. Available: https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/

[29] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 89–98.

[30] H. Liang, J. Yuan, and D. Thalman, "Egocentric hand pose estimation and distance recovery in a single rgb image," in *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, 2015, pp. 1–6.

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.

[32] M. Rad, M. Oberweger, and V. Lepetit, "Domain transfer for 3d pose estimation from color images without manual annotations," in *Proc. Asian Conf. Comput. Vis.* Springer, 2018, pp. 69–84.

[33] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3d human pose estimation in the wild: a weakly-supervised approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017.

[34] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2500–2509, 2017.

[35] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3d pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4948–4956.

[36] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman, "Single image 3d interpreter network," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 365–382.

[37] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 561–578.

[38] M. Oberweger, P. Wohlhart, and V. Lepetit, "Training a feedback loop for hand pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3316–3324.

[39] E. Dibra, T. Wolf, C. Oztireli, and M. Gross, "How to refine 3d hand pose estimation from unlabelled depth data?" in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 135–144.

[40] Y. Cai, L. Ge, J. Cai, and J. Yuan, "Weakly-supervised 3d hand pose estimation from monocular rgb images," *ECCV, Springer*, vol. 12, 2018.

[41] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang, "3d hand pose tracking and estimation using stereo matching," *Proc. Int. Conf. Image Process.*, 2017.

[42] S. Sridhar, F. Mueller, M. Zollhoefer, D. Casas, A. Oulasvirta, and C. Theobalt, "Real-time joint tracking of a hand

manipulating an object from rgb-d input," in *Proc. Eur. Conf. Comput. Vis.*, 2016. [Online]. Available: http://handtracker.mpi-inf.mpg.de/projects/RealtimeHO/

[43] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, "Real-time hand tracking under occlusion from an egocentric rgb-d sensor," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017. [Online]. Available: http://handtracker.mpi-inf.mpg.de/projects/OccludedHands/

[44] S. Lu, D. Metaxas, D. Samaras, and J. Oliensis, "Using multiple cues for hand tracking and model refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2003, pp. II–443.

[45] B. Stenger, A. Thayananthan, P. H. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical bayesian filter," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1372–1384, 2006.

[46] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2088–2095.

[47] R. Wang, S. Paris, and J. Popović, "6d hands: markerless hand-tracking for computer aided design," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 549–558.

[48] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, 2017.

[49] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3d tracking of hand articulations using kinect." in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, no. 2, 2011, p. 3.

[50] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall, "Capturing hands in action using discriminative salient points and physics simulation," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 172–193, 2016.

[51] H. Liang, J. Yuan, D. Thalmann, and Z. Zhang, "Model-based hand pose estimation via spatial-temporal hand parsing and 3d fingertip localization," *The Visual Computer*, vol. 29, no. 6-8, pp. 837–848, 2013.

[52] C. Xu and L. Cheng, "Efficient hand pose estimation from a single depth image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3456–3462.

[53] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand pointnet: 3d hand pose estimation using point sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8417–8426.

[54] L. Ge, Z. Ren, and J. Yuan, "Point-to-point regression pointnet for 3d hand pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[55] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Real-time 3d hand pose estimation with 3d convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 956–970, 2019.

[56] J. Taylor, L. Bordeaux, T. Cashman, B. Corish, C. Keskin, T. Sharp, E. Soto, D. Sweeney, J. Valentin, B. Luff *et al.*, "Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences," *ACM Trans. Graph.*, vol. 35, no. 4, p. 143, 2016.

[57] D. Tang, J. Taylor, P. Kohli, C. Keskin, T.-K. Kim, and J. Shotton, "Opening the black box: Hierarchical sampling optimization for estimating human hand pose," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3325–3333.

[58] F. Mueller, M. Davis, F. Bernard, O. Sotnychenko, M. Verschoor, M. A. Otaduy, D. Casas, and C. Theobalt, "Real-time pose and shape reconstruction of two interacting hands with a single depth camera," *ACM Trans. Graph.*, vol. 38, no. 4, p. 49, 2019.

[59] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Trans. Multimedia*, vol. 15, 2016.

[60] U. Iqbal, P. Molchanov, T. Breuel Juergen Gall, and J. Kautz, "Hand pose estimation via latent 2.5 d heatmap regression," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 118–134.

[61] P. Panteleris, I. Oikonomidis, and A. Argyros, "Using a single rgb frame for real time 3d hand pose estimation in the wild," in *International Workshop on Applications of Computer Visionn*, 2018, pp. 436–445.

[62] B. Tekin, F. Bogo, and M. Pollefeys, "H+o: Unified egocentric recognition of 3d hand-object poses and interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4511–4520.

[63] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks," *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.

[64] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11 807–11 816.

[65] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, "3d hand shape and pose estimation from a single rgb image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 833–10 842.

[66] S. Baek, K. I. Kim, and T.-K. Kim, "Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1067–1076.

[67] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 965–10 974.

[68] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "Freihand: Dataset for markerless capture of hand pose and shape from single rgb images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.

[69] S. Hampali, M. Oberweger, M. Rad, and V. Lepetit, "Ho-3d: A multi-user, multi-object dataset for joint 3d hand-object pose estimation," *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.

[70] X. Zhang, Q. Li, W. Zhang, and W. Zheng, "End-to-end hand mesh recovery from a monocular rgb image," *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.

[71] D. Kulon, H. Wang, R. A. Güler, M. Bronstein, and S. Zafeiriou, "Single image 3d hand reconstruction with mesh convolutions," *Proc. Brit. Mach. Vis. Conf.*, 2019.

[72] M. Santello, M. Flanders, and J. F. Soechting, "Postural hand synergies for tool use," *Journal of Neuroscience*, vol. 18, no. 23, pp. 10 105–10 115, 1998.

[73] E. Todorov and Z. Ghahramani, "Analysis of the synergies underlying complex hand manipulation," in *IEEE Engineering in Medicine and Biology Society*, vol. 2, 2004, pp. 4637–4640.

[74] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands deep in deep learning for hand pose estimation," *Proc. Comput. Vis. Winter Workshop*, 2015.

[75] M. Oberweger and V. Lepetit, "Deepprior++: Improving fast and accurate 3d hand pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, vol. 840, 2017, p. 2.

[76] G. Poier, D. Schinagl, and H. Bischof, "Learning pose specific representations by predicting different views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 60–69.

[77] C. Wan, T. Probst, L. Van Gool, and A. Yao, "Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[78] L. Yang and A. Yao, "Disentangling latent hands for image synthesis and pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9877–9886.

[79] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, pp. 3483–3491.

[80] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4724–4732.

[81] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 529–545.

[82] C.-H. Chen and D. Ramanan, "3d human pose estimation= 2d pose estimation+ matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, no. 5, 2017, p. 6.

[83] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[84] M. Oberweger, P. Wohlhart, and V. Lepetit, "Generalized feedback loop for joint hand-object pose estimation." *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[85] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4068–4076.

[86] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2016, pp. 424–432.

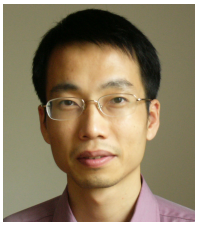
[87] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learning*, 2015, pp. 448–456.

[88] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.

[89] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2017.
[90] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and robust hand tracking from depth," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1106–1113.

Yujun Cai received the B.Eng. degree in Information Science and Engineering from Southeast University in 2017. She is now a PhD candidate with the Institute for Media Innovation, Interdisciplinary Graduate School, Nanyang Technological University, Singapore. Her research interests mainly include computer vision, machine learning and human-computer interaction.

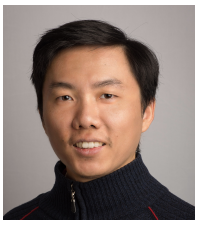
Liuhao Ge received the B.Eng. degree in Detection Guidance and Control Technology from Nanjing University of Aeronautics and Astronautics in 2011, the M.Eng. degree in Control Theory and Engineering from Southeast University in 2014, and the PhD degree from Nanyang Technological University in 2019. His research interests mainly include computer vision, machine learning, and human-computer interaction.



JianfeiCai (S'98-M'02-SM'07) received his PhD degree from the University of Missouri-Columbia. He is currently a Professor and serves as the Head of the Data Science & AI Department at Faculty of IT, Monash University, Australia. Before that, he had served as a Cluster Deputy Director of Data Science & AI Research Center (DSAIR), Head of Visual and Interactive Computing Division and Head of Computer Communications Division in Nanyang Technological University (NTU). His major research interests include visual computing, computer vision and multimedia. He has published over 200 technical papers in international journals and conferences. He has served as an Associate Editor for IEEET-IP, T-MM, T-CSVT and Visual Computer as well as serving as Area Chair for ICCV, ECCV, ACM Multimedia, ICME and ICIP. He was the Chair of IEEE CAS VSPC-TC during 2016-2018. He had also served as the leading TPC Chair for IEEE ICME 2012.

Nadia Magnenat Thalmann is Professor and Director of the Institute for Media Innovation in NTU, Singapore. She is also the Director of MIRALab at the University of Geneva in Switzerland, a ground-breaking interdisciplinary research institute that has published more than 700 papers in top journals or conferences. Her research domains are the simulation of Virtual Humans and Social Robots, including the recognition of faces, emotions and gestures. In NTU, Singapore, she revolutionized social robotics by unveiling the first realistic social robot Nadine that can show mood and emotions and remember people and actions.

Besides having bachelor's and master's degrees in disciplines such as psychology, biology, chemistry and computer science, Professor Thalmann completed her PhD in quantum physics at the University of Geneva. She has received honorary doctorates from Leibniz University of Hannover and the University of Ottawa in Canada and several prestigious other Awards as the Humboldt Research Award in Germany. She is a life member of the Swiss Academy of Technical Sciences.



Junsong Yuan (M'08' SM'14) is currently an Associate Professor and Director of Visual Computing Lab at Department of Computer Science and Engineering (CSE), State University of New York at Buffalo, USA. Before that he was an Associate Professor at Nanyang Technological University (NTU), Singapore. He obtained his Ph.D. from Northwestern University. He received Best Paper Award from IEEE Trans. on Multimedia, Nanyang Assistant Professorship from NTU, and Outstanding EECS Ph.D. Thesis award from Northwestern University. He is currently Senior Area Editor of Journal of Visual Communications and Image Representation (JVCI), Associate Editor of IEEE Trans. on Image Processing (T-IP) and Machine Vision and Applications (MVA). He is Program Co-Chair of IEEE Conf. on Multimedia Expo (ICME'18) and Steering Committee Member of ICME (2018-2019). He also served as Area Chair for CVPR, ICIP, ICPR, ACCV, ACM MM, WACV etc. He is a Fellow of International Association of Pattern Recognition (IAPR).