Weakly-supervised 3D Hand Pose Estimation from Monocular RGB Images *

Yujun Cai[1][0000−0002−0993−4024], Liuhao Ge[1], Jianfei Cai[2], and Junsong Yuan[3]

[1] Institute for Media Innovation, Interdisciplinary Graduate School, Nanyang Technological University
{yujun001,ge0001ao}@e.ntu.edu.sg
[2] School of Computer Science and Engineering, Nanyang Technological University
asjfcai@ntu.edu.sg
[3] Department of Computer Science and Engineering, State University of New York at Buffalo University
jsyuan@buffalo.edu

Abstract. Compared with depth-based 3D hand pose estimation, it is more challenging to infer 3D hand pose from monocular RGB images, due to substantial depth ambiguity and the difficulty of obtaining fully-annotated training data. Different from 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 dataset with the aid of a depth regularizer, which generates depth maps from predicted 3D pose and serves as weak supervision for 3D pose regression. Extensive experiments on benchmark datasets validate the effectiveness of the proposed depth regularizer in both weakly-supervised and fully-supervised settings.

Keywords: 3D hand pose estimation, weakly-supervised methods, depth regularizer

1 Introduction

Articulated hand pose estimation has aroused a long-standing study in the past decades [23, 38, 39], since it plays a significant role in numerous applications

* This research is supported by the BeingTogether Centre, a collaboration between Nanyang Technological University (NTU) Singapore and University of North Carolina (UNC) at Chapel Hill. The BeingTogether Centre is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative. This research is also supported in part by Singapore MoE Tier-2 Grant (MOE2016-T2-2-065) and start-up funds from University at Buffalo.

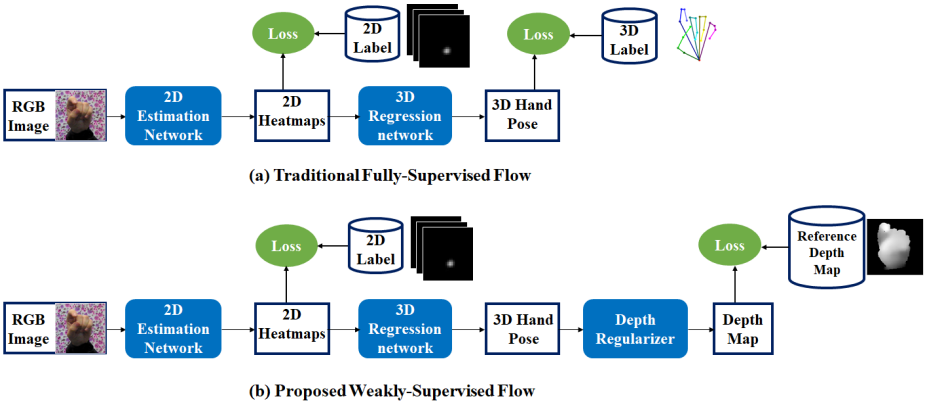

Fig. 1. Illustration of the concept of weakly supervised 3D hand pose estimation. Different from conventional fully-supervised methods (a) that use 3D labels to guide joint predictions, our proposed weakly-supervised method (b) leverages the reference depth map, which can be easily obtained by consumer-grade depth camera, to provide weak supervision. 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.

such as human-computer interaction and virtual reality. Although 3D hand pose estimation with depth cameras [13, 7, 26, 41, 6] has gained tremendous success in recent years, the advance in monocular RGB-based 3D hand pose estimation[46, 18, 27, 15], however, still remains limited. Due to the wide availability of RGB cameras, the RGB-based solution for 3D hand pose estimation is more favored than depth-based solutions in many vision applications.

Compared with depth images, single-view RGB images exhibit inherent depth ambiguity, which makes 3D hand pose estimation from single RGB images a challenging problem. To overcome the ambiguity, recent work on RGB-based 3D hand pose estimation [46] relies 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 annotator. To address this problem, Zimmermann *et al.* [46] 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, however, does not generalize well to real image dataset due to the domain shift between image features. Paschalis[22] employed a discriminative approach to localize the 2D keypoints and model fitting method to calculate the 3D pose. Recently, Muller *et al.* [18] leveraged CycleGANs [45] to generate a "real" dataset

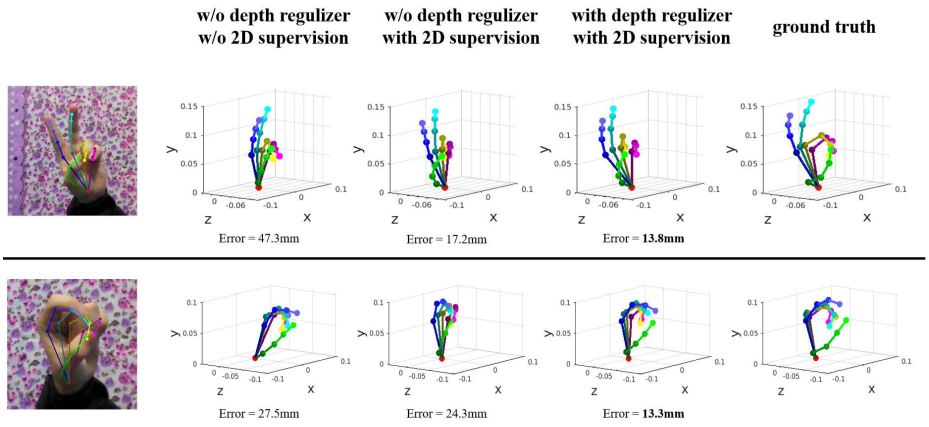

Fig. 2. We present a weakly-supervised approach for 3D hand pose estimation from monocular RGB-only input. Our method with depth regularizer (column 4) significantly boosts the performance of other baselines (column 2 and column 3). Note that columns 2-5 are shown in a novel viewpoint for better comparison.

transferred from synthetic dataset. However, limited performance shows that there still exists gap between generated "real" images and real-world images.

Our proposed weakly-supervised adaptation method addresses this limitation in a novel perspective. We observe that most of the previous works [46, 18, 27] 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. Moreover, it is indicated that such low-cost depth images contain rich cues for 3D hand pose labels, as depth-based methods show decent 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. 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 [44, 32, 42, 37, 1] in body pose estimation, we apply a cascaded network architecture including a 2D pose estimation network and a 3D regression network. We note that 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 [19, 4], we innovate the structure with a depth regularizer, which generates depth images from predicted 3D hand pose and regularizes the predicted 3D regression by supervising the rendered depth map, as shown in Figure 1 (b). This network essentially learns the mapping from 3D pose to its corresponding depth map, which can be used for the knowledge transfer from

the fully-annotated synthetic images to weakly-labeled real-world images without entire 3D annotations. Additionally, we apply the depth regularizer to the fully-supervised setting. The effectiveness of the depth regularizer is experimentally verified for both our weakly-supervised and fully-supervised methods on two benchmark datasets: RHD[46] and STB datasets[43].

To summarize, this work makes the following contributions:

- We innovatively introduce the weakly supervised problem of leveraging low-cost depth maps during training for 3D hand pose estimation from RGB images, which releases the burden of 3D joint labeling.
- We propose an end-to-end learning based 3D hand pose estimation model for weakly-supervised adaptation from fully-annotated synthetic images to weakly-labeled real-world images. Particulary, we introduce a depth regularizer supervised by the easily captured depth images, which considerably enhances the estimation accuracy compared with weakly-supervised baselines (see Figure 2).
- We conduct experiments on the two benchmark datasets, which show that our weakly-supervised approach compares favorably with existing works and our proposed fully-supervised method outperforms all the state-of-the-art methods.

2 Related Work

3D hand pose estimation has been studied extensively for a long time, with vast theoretical innovations and important applications. Early works [23, 17, 28] 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 [21, 35] 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[10], 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 [20, 34, 16], discriminative approaches [13, 40, 6, 7, 5, 8] and hybrid approaches [25, 31, 30].

Inspired by the great improvement of CNN-based 3D hand pose estimation from depth images[24], deep learning has also been adopted in some recent works on monocular RGB-based applications [46, 18]. In particular, Zimmermann *et al.* [46] proposed a deep network that learns an implicit 3D articulation prior of joint locations in canonical coordinates, as well as constructs a synthetic dataset to

tackle the problem of insufficient annotations. Muller *et al.* [18] 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 weakly-labeled real-world images.

Dibra *et al.* [4] is the closest work in spirit to our approach, which proposed 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 [4] in several aspects. Firstly, our work is targeted at 3D hand pose estimation from single RGB input, whereas [4] focuses on depth-based predictions. Secondly, compared with [4] that leverages a rigged 3D hand model to synthesize depth images, we use a simple fully-convolutional network to infer the corresponding depth maps from the predicted 3D hand pose. To the best of our knowledge, 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 an alternative solution for this problem and will enable further research of utilizing depth images in RGB-input applications.

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 3 depicts the proposed network architecture, which utilizes a cascaded architecture inspired from [44]. It consists of a 2D pose estimation network (convolutional pose machines - CPM), a 3D regression network, and a depth regularizer. Given a cropped single RGB image containing human hand with certain gesture, we aim to get the 2D heatmap and the corresponding depth of each joint from the proposed end-to-end network. The 2D joint locations are denoted as $\mathbf{\Phi}_{2D} \in \mathbf{\Lambda}_{2D}$, where $\mathbf{\Lambda}_{2D} \in \mathcal{R}^{K\times 2}$ and the depth values are denoted as $\mathbf{\Phi}_z \in \mathbf{\Lambda}_z$, where $\mathbf{\Lambda}_z \in \mathcal{R}^{K\times 1}$. The final output 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. Note that our depth regularizer is only utilized during training. During testing, only 2D estimation network and regression network are used to predict joint locations.

The depth regularizer is the key part to facilitate the proposed weakly supervised training, *i.e.*, relieve the painful joint depth annotations for real-world dataset by making use of the rough depth maps, which can be easily captured by consumer-grade depth cameras. In addition, our experiments show that the introduced depth regularizer can slightly improve 3D hand pose prediction of

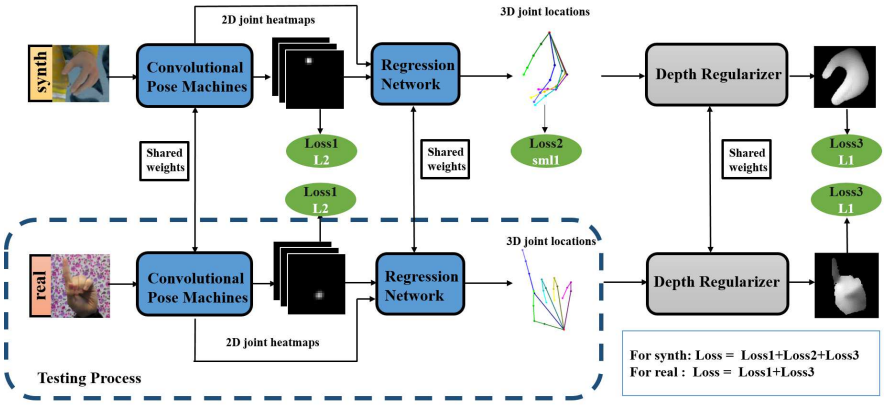

Fig. 3. Overview of our proposed weakly-supervised 3D hand pose regression network, which is trained in an end-to-end manner. During training, cropped images from both synthetic dataset and real image dataset are mixed in each single batch as the input to the network. To compensate the absence of ground truth annotations for joint depth in real data, we extend the network with a depth regularizer by leveraging the corresponding depth maps available in both synthetic and real datasets to provide a weak supervision. During testing, real images only go through the part of the network in the dashed line box. The obtained 2D heatmaps and joint depth are concatenated as the output of the network.

fully-supervised methods as well, since it serves as an additional constraint for the 3D hand pose space.

The entire network is trained with a Rendered Hand Pose Dataset (RHD) created by [46] and a real-world dataset from Stereo Hand Pose Tracking Benchmark [43]. 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 pretrained on I_{RHD} and then adapted to I_{STB} by fusing the training of both datasets. For fully-supervised learning, the two datasets are used independently in the training and evaluation processes.

3.2 2D Pose Estimation Network

For 2D pose estimation, we adopt the encoder-decoder architecture similar to the Convolutional Pose Machines by Wei *et al.* [36] and [46], 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 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 probabilites in the heat map, as proposed in [29]. We initialize the network with

weights adapted from human pose prediction to I_{RHD}, tuned by Zimmermann *et al.* [46].

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

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

3.3 Regression Network

The objective of the regression network is to infer the depth of each joint from the obtained 2D heatmap. Most previous work [46, 2, 32] in 3D human pose and hand pose estimation based on single image attempt to lift the set of 2D heatmaps into 3D space dirctly, while a key issue for this strategy is how to distinguish between the multiple 3D poses inferred from a single 2D skeleton. Inspired from [44], our method exploits contextual information to reduce the ambiguity of lifting 2D heatmaps to 3D locations, by extracting the intermediate image evidence in 2D pose estimation network concatenated with the predicted 2D heatmaps as the input to the regression network. We employ a simple yet effective depth regression network structure with only two convolutional layers and three fully-connected layers. Note that here we infer a scale-invariant and translation-invariant representation of joint depth, by subtracting each hand joint with the location of root keypoint and then normalizing it by the distance between a certain pair of keypoints, as done in [46, 18].

For fully-supervised learning, we simply apply smooth L1 loss introduced in [9] between our predicted joint depth $\hat{\mathbf{\Phi}}_z$ and the ground truth label $\mathbf{\Phi}_z^{gt}$. For weakly-supervised learning, no penalty is enforced because of the absence of 3D annotations. To address this issue, we introduce a novel depth regularizer as weak supervision for joint depth regression, which will be elaborated in Section 3.4.

Overall, the loss function of the regression network is defined as

$$L_z(\hat{\mathbf{\Phi}}_z, \mathbf{\Phi}_z^{gt}) = \begin{cases} smooth_{L1}(\hat{\mathbf{\Phi}}_z, \mathbf{\Phi}_z^{gt}) & , if \text{ full supervision} \\ 0 & , if \text{ weak supervision} \end{cases} \tag{2}$$

in which

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, & if|x| < 1 \\ |x| - 0.5, & otherwise. \end{cases} \tag{3}$$

3.4 Depth Regularizer

The purpose of the depth regularizer is to take the easily-captured depth images as an implicit constraint of physical structures that can be applied to both weakly-supervised and fully-supervised situations. Figure 4 shows the architecture of the proposed depth regularizer, which is fully-convolutional with six layers, inspired by [19, 3]. Each layer contains a transposed convolution followed by

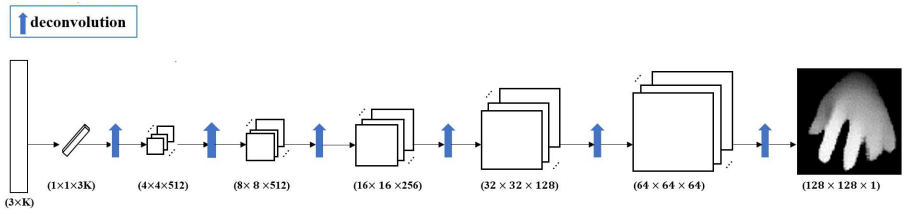

Fig. 4. Network architecture of our proposed depth regularizer. Given 3D hand joint locations as the input, the depth regularizer is able to render the corresponding depth map by gradually enlarging the intermediate feature maps and finally combining them into a single depth image.

a Relu, after which the feature map is expanded along both image dimensions. In the first five layers, batch normalization [12] and drop out [11] are introduced before Relu in order to reduce the dependency on the initialization and alleviate from overfitting the training data. The final layer combines all feature maps to generate the corresponding depth image from 3D hand pose.

Let $(\hat{\mathbf{\Phi}}_{3D}, D)$ denote a training sample, where $\hat{\mathbf{\Phi}}_{3D}$ is the input of the depth regularizer containing a set of 3D hand joint locations, and $\mathbf{D}$ is the corresponding depth image. We normalize $\mathbf{D}$ into $\mathbf{D}_n$:

$$\mathbf{D}_n = \sum_{i,j} \frac{d_{max} - d_{ij}}{d_{range}} \tag{4}$$

where d_{ij} is the depth value at the image location (i, j), and d_{max} and d_{range} represent the maximum depth value and the depth range, respectively. Note that the normalized depth value tends to be larger when located closer to the camera and background is set to 0 in this process.

The input of the network $\hat{\mathbf{\Phi}}_{3D} = \{(\mathbf{\Phi}^{gt}_{2D}, \mathbf{X}_z)\}$ contains two parts: the ground truth 2D labels $\mathbf{\Phi}^{gt}_{2D}$ in the image coordinate system and the joint depth $\mathbf{X}_z$. Note that the reason we use ground truth 2D locations rather than our predicted 2D results is to simplify the training process since no back-propagation from the depth regularizer is fed back into the 2D pose estimation network. For the joint depth $\mathbf{X}_z$, we apply the same normalization:

$$\mathbf{X}_z = \frac{d_{max} - \hat{\mathbf{\Phi}}_z \cdot L_{scale} - d_{root}}{d_{range}} \tag{5}$$

where $\hat{\mathbf{\Phi}}_z$ denotes the predicted joint depth from the regression network, which is a set of root-relative and normalized values and can be recovered to global coordinates by multiplying with hand scale L_{scale} and shifting to root depth d_{root}.

To train the depth regularizer, we adopt L1 norm to minimize the difference between the generated depth image $\hat{\mathbf{D}}_n$and the corresponding ground truth $\mathbf{D}_n$:

$$L_{dep}(\hat{\mathbf{D}}_n, \mathbf{D}_n) = |\hat{\mathbf{D}}_n - \mathbf{D}_n| \tag{6}$$

3.5 Training

Combining the losses in Eq. (1), (2), and (6), we obtain the overall loss function as

$$L = \lambda_{2D} L_{2D}(\hat{\mathbf{\Phi}}_{HM}, \mathbf{\Phi}_{2D}^{gt}) + \lambda_z L_z(\hat{\mathbf{\Phi}}_z, \mathbf{\Phi}_z^{gt}) + \lambda_{dep} L_{dep}(\hat{\mathbf{D}}_n, \mathbf{D}_n). \quad (7)$$

Adam optimization [14] is used for training. For weakly-supervised learning, similar to [44] and [33], we adopt fused training where each mini-batch contains both the synthetic and the real training examples (half-half), shuffled randomly during the training process. 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* initializes the regression network and fine-tunes the 2D pose estimation network with weights from Zimmermann *et al.* [46], which are adapted from the Convolutional Pose Machines [36]. *Stage 2* initializes the depth regularizer, as described in Section 3.4. *Stage 3* fine-tunes the whole network with all the training data, which is an end-to-end training.

4 Experiments

4.1 Implementation Details

Our method is implemented with Pytorch. For the first training stage described in Section 3.5, we take 60 epochs with an initial learning rate of 10^{-7}, a batch size of 8 and a regularization strength of 5×10^{-4}. For Stage 2 and Stage 3, we spend 40 and 20 epochs, respectively. During the fine-tunning process of the whole network, we set $\lambda_{2D} = 1$, $\lambda_z = 0.1$ and $\lambda_{dep} = 1$. All experiments are conducted on one GeForce GTX 1080 GPU with CUDA 8.0.

4.2 Datasets and Metrics

We evaluate our method on two publicly available datasets: Rendered Hand Pose Dataset (RHD) [46] and a real-world dataset from Stereo Hand Pose Tracking Benchmark (STB) [43].

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 regression 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

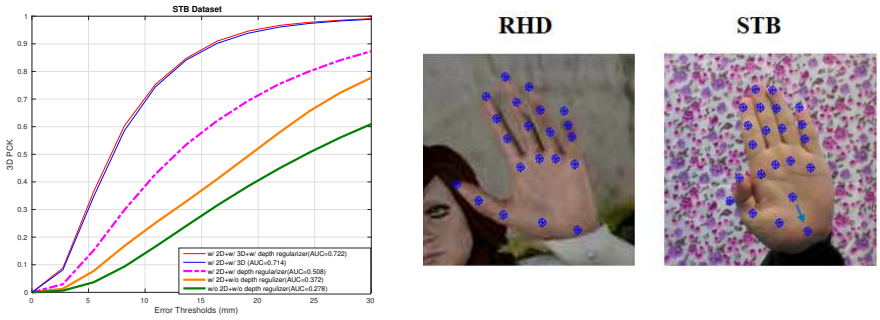

Fig. 5. Left: Comparisons of 3D PCK results of different baselines with our method on STB [43]. Our proposed weakly-supervised method, w/ 2D + w/ depth regularizer, significantly outperforms other weakly-supervised baselines(Orange and Green curve). Right: Different annotation schemes on RHD [46] and STB [43] dataset. Note that we move the root joint location of STB dataset from palm to wrist keypoint to make the two datasets consistent with each other.

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 with 2D annotations, as well as root depth (*i.e.*, wrist 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 [46, 18], which trains on 10 sequences and tests on the other two.

We evaluate the 3D hand pose estimation performance with two 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 [46, 18]. The second metric is the mean error distance in z-dimension over all testing frames, which is used to further analyse the impact of the proposed depth regularizer. Following the same condition used in [46, 18], we assume that the global hand scale and the root depth are known in the experimental evaluations so that we can report PCK curve based on 3D hand joint locations in the global domain, which are computed from the output root-relative articulations.

4.3 Quantitative Results

Weak supervision. We first evaluate the impact of weak label constraints on STB dataset compared with fully-supervised methods with complete 2D and 3D

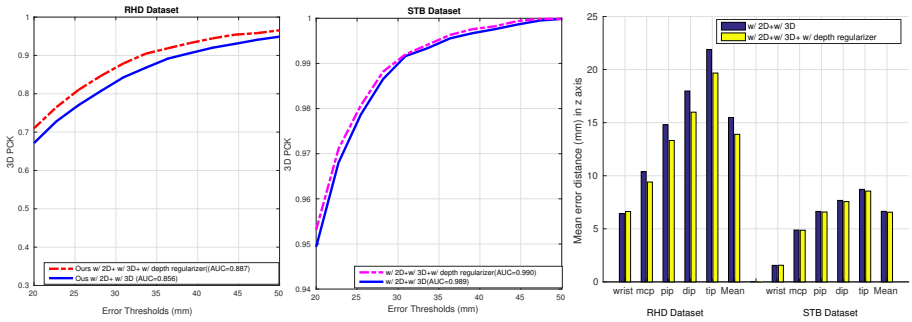

Fig. 6. The effect of the proposed depth regularizer in fully-supervised setting on RHD [46] and STB datasets [43]. Left: 3D PCK on RHD dataset. Middle: 3D PCK on STB dataset. Right: mean joint error distances in z-dimension on RHD and STB datasets.

annotations. Specifically, we compare our proposed weakly-supervised approach (**w/ 2D + w/ depth regularizer**) with three baselines: a) **w/o 2D + w/o depth regularizer**: directly using pretrained model based on RHD dataset; b) **w/ 2D + w/o depth regularizer**: tuning the pretrained network with 2D labels in STB dataset and c) **w/ 2D + w/ 3D**: fully-supervised method with complete 2D and 3D annotations.

As illustrated in the left part of Figure 5, the fully-supervised method 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. Compared with baseline-a, by fine-tuning the pretrained model with the 2D labels of the real images, baseline-b significantly improves the AUC value from 0.667 to 0.807. Moreover, adding our proposed depth regularizer further increases AUC to 0.889, which demonstrates the effectiveness of the depth regularizer.

We note that STB and RHD datasets adopt different schemes for 2D and 3D annotations, as shown in the right part of Figure 5. In particular, STB dataset annotates palm position as root joint, which is different from RHD dataset that uses wrist position as root keypoint. Thus, we move the palm joint in STB to wrist point so as to make the annotations consistent for fused training. To evaluate the introduced noise of moving root joint, we compare our results of fully-supervised method on STB dataset with palm-relative and wrist-relative representations. Original palm-relative representation performs slightly better, reducing the mean error by about 0.6mm. Besides, it is also noted that MCP (Metacarpophalangeal joints) positions are closer to wrist joint in STB dataset and labels for STB dataset are relatively noisy compared with synthetic dataset

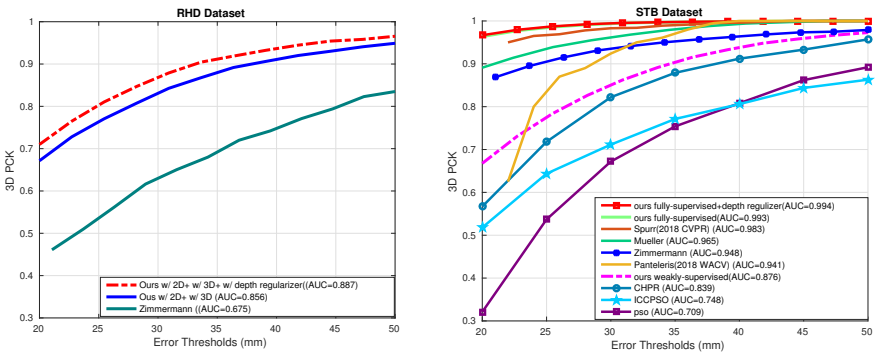

Fig. 7. Comparisons with state-of-the-art methods on RHD [46] and STB [43]. Left: 3D PCK on RHD dataset. Right: 3D PCK on STB dataset.

RHD (*e.g.*, thumb dip is annotated in the background). Due to these differences, we argue that there exists a bias between our pose predictions and the ground truth provided by STB dataset, which might decrease the reported estimation accuracy of our proposed weakly-supervised approach. Furthermore, these inconsistencies, on the other hand, suggest the necessity of the introduced depth regularizer, since it provides certain prior knowledge of hand pose and shapes.

Fully-supervised 3D Hand Pose Estimation. We also evaluate the effectiveness of the depth regularizer in the fully-supervised setting on both RHD and STB datasets. Note that the two datasets are trained independently in this case. As presented in Figure 6 (left) and Figure 6 (middle), our fully-supervised method with depth regularizer outperforms that without depth regularizer on both RHD and STB dataset, with improvement of 0.031 and 0.001 in AUC, respectively. Figure 6 (right) shows the mean joint error in z-dimension, indicating that adding depth regularizer is able to slightly improve the fully-supervised results in the joint depth estimation.

Comparisons with State-of-the-arts. Fig 7 shows the comparisons with state-of-the-art methods [46, 43, 18, 27, 22] on both RHD and STB datasets. It can be seen that on RHD dataset, even without the depth regularizer, our fully-supervised method significantly outperforms the state-of-the-art method [46], improving the AUC value from 0.675 to 0.887. On STB dataset, our fully-supervised method achieves the best results compared with all existing methods. Note that our weakly-supervised method is also superior to some of the existing works, which demonstrates the potential values for the weakly-supervised exploration when complete 3D annotations are difficult to obtain in real-world dataset. It is also noted that the AUC values of our proposed methods in Figure 7 are slightly different from their counterparts in Section 4.3. This is because here

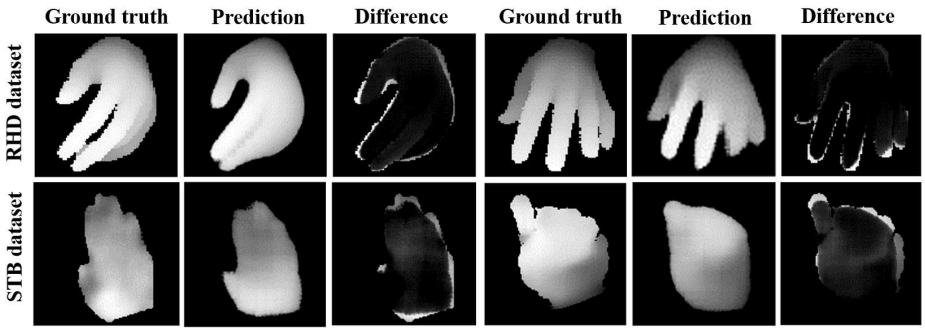

Fig. 8. Samples of the generated depth maps by the trained depth regularizer with the input of ground truth 3D hand joint locations. Our trained depth regularizer is able to render plausible and convincing depth maps. Note that the errors are mainly located around contours of the hand, where the reference depth images (e.g. captured by depth camera) are typically noisy.

we test on the stereo pair subset STB-BB rather than the color-depth subset STB-SK.

4.4 Qualitative Results

Figure 9 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, after additionally imposing the depth regularizer with the reference depth images, our weakly-supervised approach on real-world dataset yields considerably better estimation accuracy, especially in terms of global orientation, which is consistent with our aforementioned quantitative analysis.

Figure 10 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 with the depth regularizer is robust to various hand orientations and complicated pose articulations.

Although the depth regularizer is only used in training but not in testing, it is interesting to see whether it has learned a manifold of hand poses. Thus, we collect some samples of the depth images generated by our well trained depth regularizer, given ground truth 3D hand joint locations, as shown in Figure 8. We can see that our depth regularizer is able to to render smooth and convincing depth images for hand poses in large variations and self-occlusions.

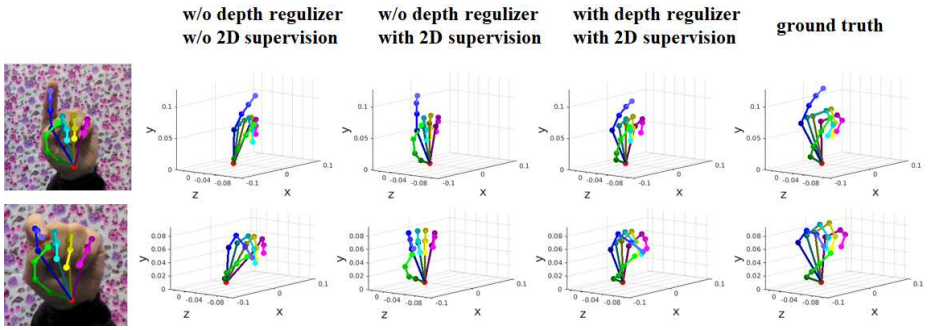

Fig. 9. Visual results of our proposed weakly-supervised approach (column 1,4) and other baselines (column 2,3), compared with ground truth (column 5). Note that columns 2-5 are shown at a novel viewpoint for easy comparison.

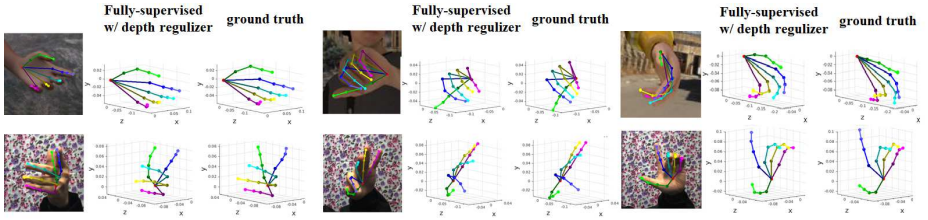

Fig. 10. Visual results of our fully-supervised method on RHD and STB datasets. First row: RHD dataset. Second row: STB dataset. Note that skeletons are shown at a novel viewpoint for easy comparison.

5 Conclusions

Building a large real-world hand dataset with full 3D annotations is often one of the major bottlenecks for learning-based approaches in 3D hand pose estimation task. To address this problem, our approach presents one way to adapt weakly-labeled real-world dataset from 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. To be specific, we introduce a simple yet effective end-to-end architecture consisting of a 2D estimation network, a regression network and a novel depth regularizer. Quantitative and qualitative experimental results show that our weakly-supervised method compares favorably with the existing works and our fully-supervised approach considerably outperforms the state-of-the-art methods. We note that we only show one way for weakly-supervised 3D hand pose estimation. There is a large space for un-/weakly-supervised learning.

References

1. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision. pp. 561–578. Springer (2016)
2. Chen, C.H., Ramanan, D.: 3d human pose estimation= 2d pose estimation+ matching. In: CVPR. vol. 2, p. 6 (2017)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 424–432. Springer (2016)
4. Dibra, E., Wolf, T., Oztireli, C., Gross, M.: How to refine 3d hand pose estimation from unlabelled depth data? In: 3D Vision (3DV), 2017 International Conference on. pp. 135–144. IEEE (2017)
5. Ge, L., Cai, Y., Weng, J., Yuan, J.: Hand pointnet: 3d hand pose estimation using point sets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8417–8426 (2018)
6. Ge, L., Liang, H., Yuan, J., Thalmann, D.: Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3593–3601 (2016)
7. Ge, L., Liang, H., Yuan, J., Thalmann, D.: 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, p. 5 (2017)
8. Ge, L., Ren, Z., Yuan, J.: Point-to-point regression pointnet for 3d hand pose estimation. In: Proc. European Conf. Comput. Vis. (2018)
9. Girshick, R.: Fast r-cnn. In: Computer Vision (ICCV), 2015 IEEE International Conference on. pp. 1440–1448. IEEE (2015)
10. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al.: Recent advances in convolutional neural networks. Pattern Recognition (2017)
11. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
12. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456 (2015)
13. Keskin, C., Kıraç, F., Kara, Y.E., Akarun, L.: Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In: European Conference on Computer Vision. pp. 852–863. Springer (2012)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Liang, H., Yuan, J., Thalman, D.: Egocentric hand pose estimation and distance recovery in a single rgb image. In: Multimedia and Expo (ICME), 2015 IEEE International Conference on. pp. 1–6. IEEE (2015)
16. Liang, H., Yuan, J., Thalmann, D., Zhang, Z.: Model-based hand pose estimation via spatial-temporal hand parsing and 3d fingertip localization. The Visual Computer **29**(6-8), 837–848 (2013)

17. Lu, S., Metaxas, D., Samaras, D., Oliensis, J.: Using multiple cues for hand tracking and model refinement. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. vol. 2, pp. II–443. IEEE (2003)
18. 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: Proceedings of Computer Vision and Pattern Recognition (CVPR) (June 2018), https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/
19. Oberweger, M., Wohlhart, P., Lepetit, V.: Training a feedback loop for hand pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3316–3324 (2015)
20. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Efficient model-based 3d tracking of hand articulations using kinect. In: BmVC. vol. 1, p. 3 (2011)
21. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 2088–2095. IEEE (2011)
22. Panteleris, P., Oikonomidis, I., Argyros, A.: Using a single rgb frame for real time 3d hand pose estimation in the wild. In: Applications of Computer Vision (WACV), 2018 IEEE Winter Conference on. pp. 436–445. IEEE (2018)
23. Rehg, J.M., Kanade, T.: Digiteyes: Vision-based hand tracking for human-computer interaction. In: Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on. pp. 16–22. IEEE (1994)
24. Ren, Z., Yuan, J., Meng, J., Zhang, Z.: Robust part-based hand gesture recognition using kinect sensor. IEEE Trasactions on Multimedia **15** (2016)
25. Sharp, T., Keskin, C., Robertson, D., Taylor, J., Shotton, J., Kim, D., Rhemann, C., Leichter, I., Vinnikov, A., Wei, Y., et al.: Accurate, robust, and flexible real-time hand tracking. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 3633–3642. ACM (2015)
26. Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., et al.: Efficient human pose estimation from single depth images. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(12), 2821–2840 (2013)
27. Spurr, A., Song, J., Park, S., Hilliges, O.: Cross-modal deep variational hand pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 89–98 (2018)
28. Stenger, B., Thayananthan, A., Torr, P.H., Cipolla, R.: Model-based hand tracking using a hierarchical bayesian filter. IEEE transactions on pattern analysis and machine intelligence **28**(9), 1372–1384 (2006)
29. Sun, X., Xiao, B., Liang, S., Wei, Y.: Integral human pose regression. arXiv preprint arXiv:1711.08229 (2017)
30. Tang, D., Taylor, J., Kohli, P., Keskin, C., Kim, T.K., Shotton, J.: Opening the black box: Hierarchical sampling optimization for estimating human hand pose. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3325–3333 (2015)
31. Taylor, J., Bordeaux, L., Cashman, T., Corish, B., Keskin, C., Sharp, T., Soto, E., Sweeney, D., Valentin, J., Luff, B., et al.: Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. ACM Transactions on Graphics (TOG) **35**(4), 143 (2016)
32. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. CVPR 2017 Proceedings pp. 2500–2509 (2017)

33. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Computer Vision (ICCV), 2015 IEEE International Conference on. pp. 4068–4076. IEEE (2015)
34. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. International Journal of Computer Vision **118**(2), 172–193 (2016)
35. Wang, R., Paris, S., Popović, J.: 6d hands: markerless hand-tracking for computer aided design. In: Proceedings of the 24th annual ACM symposium on User interface software and technology. pp. 549–558. ACM (2011)
36. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4724–4732 (2016)
37. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single image 3d interpreter network. In: European Conference on Computer Vision. pp. 365–382. Springer (2016)
38. Wu, Y., Huang, T.S.: Capturing articulated human hand motion: A divide-and-conquer approach. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. vol. 1, pp. 606–611. IEEE (1999)
39. Wu, Y., Huang, T.S.: View-independent recognition of hand postures. In: cvpr. p. 2088. IEEE (2000)
40. Xu, C., Cheng, L.: Efficient hand pose estimation from a single depth image. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 3456–3462. IEEE (2013)
41. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 1385–1392. IEEE (2011)
42. Yasin, H., Iqbal, U., Kruger, B., Weber, A., Gall, J.: A dual-source approach for 3d pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4948–4956 (2016)
43. Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., Yang, Q.: 3d hand pose tracking and estimation using stereo matching. arXiv preprint arXiv:1610.07214 (2016)
44. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: IEEE International Conference on Computer Vision (2017)
45. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on. pp. 2242–2251. IEEE (2017)
46. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: International Conference on Computer Vision (2017)