Weakly-supervised Domain Adaptation via GAN and Mesh Model for Estimating 3D Hand Poses Interacting Objects

Seungryul Baek
Imperial College London
s.baek15@imperial.ac.uk

Kwang In Kim
UNIST
kimki@unist.ac.kr

Tae-Kyun Kim
Imperial College London
tk.kim@imperial.ac.uk

Abstract

Despite recent successes in hand pose estimation, there yet remain challenges on RGB-based 3D hand pose estimation (HPE) under hand-object interaction (HOI) scenarios where severe occlusions and cluttered backgrounds exhibit. Recent RGB HOI benchmarks have been collected either in real or synthetic domain, however, the size of datasets is far from enough to deal with diverse objects combined with hand poses, and 3D pose annotations of real samples are lacking, especially for occluded cases. In this work, we propose a novel end-to-end trainable pipeline that adapts the hand-object domain to the single hand-only domain, while learning for HPE. The domain adaption occurs in image space via 2D pixel-level guidance by Generative Adversarial Network (GAN) and 3D mesh guidance by mesh renderer (MR). Via the domain adaption in image space, not only 3D HPE accuracy is improved, but also HOI input images are translated to segmented and de-occluded hand-only images. The proposed method takes advantages of both the guidances: GAN accurately aligns hands, while MR effectively fills in occluded pixels. The experiments using Dexter-Object, Ego-Dexter and HO3D datasets show that our method significantly outperforms state-of-the-arts trained by hand-only data and is comparable to those supervised by HOI data. Note our method is trained primarily by hand-only images with pose labels, and HOI images without pose labels.

1. Introduction

Estimating 3D hand poses either from RGB images [59, 86, 7, 22, 35, 61] or depth maps [83, 38, 66, 75, 38, 81, 34, 50, 81, 1] has shown great improvements [68, 59, 75, 83, 14, 79, 51, 9, 17, 80, 62, 77, 48, 84, 72, 69, 13, 60, 39, 65, 40, 27, 33, 42, 12, 81]. The attributes behind successful hand pose estimation are: deep learning methods that are able to learn highly non-linear 2D-to-3D mapping, and available large-scale datasets [83, 59] which enable sufficient training of convolutional neural networks (CNNs). However, challenges are still remaining for hand-object interaction (HOI) cases where there are severe occlusions and cluttered backgrounds.

(a) (b) (c) (d) (e)

Figure 1: Example hand pose estimation results in the hand-object interaction scenario. Our method synthesizes and gradually refines the hand-only images from the input HOI images. (a) input images $\mathbf{x}$; (b) initial hand-only estimates $\mathbf{x}'$ constructed by our mesh renderer based on $\mathbf{x}$; (c) refinements $\mathbf{x}''$ of $\mathbf{x}'$ generated by GAN; (d-e) final hand-only images $\mathbf{z}$ and skeleton estimates $\mathbf{y}$ generated based on $\mathbf{x}''$.

RGB benchmarks have been recently proposed to tackle the scenario [62, 35, 21, 20, 11], where CNN-based hand pose estimator (HPE)s are trained using HOI images. Collecting quality 3D pose annotations of real RGB images, however, is challenging due to e.g. occlusions under the HOI scenarios. A complete and automatic pipeline for annotating 3D joint locations for severely occluded hands does not exist. It either requires much manual effort to continuously check and refine the labels [87] or the use of magnetic sensors [12]/data gloves [5] corrupts RGB images. Alternatively, they resort to synthetic data. Most existing large-scale datasets for hand pose estimation (e.g. *RHD* [86], SynthHands [37], GANerated [35] and *Obman* [21]) are synthetic. Real datasets have either limited annotations such as discrete grasp types (e.g. GUN-71 [52]), only 5 finger

tips (e.g. *DO* [62], EGO [37]) or a limited number of frames (e.g. *HO3D* [20]). FPHA [12] dataset is real and fair-sized; however their RGB frames are corrupted since the magnetic sensors used are visible. FreiHand [87] is of the latest benchmarks having a moderate-scale (35k). However, only less than half of it contains HOI images. When considering diverse objects, backgrounds and the large hand pose space, far more samples are required for training. In [21], the authors reported the accuracy of hand pose estimator trained and tested using either 'hand-only' or 'HOI' data. When the hand pose estimator is trained by 'HOI', it does not perform well on hand-only testing images in comparison to the model trained by 'hand-only', while increasing the accuracy on HOI testing images.

In this paper, we aim at adapting the domain of hand-object interaction (HOI) to the domain of single hand-only (See Fig. 1 for examples). This helps reduce the number of 3D joint annotations of HOI samples to train HPE. To the best of our knowledge, this is the first work to fill the gap between the HOI and hand-only domains. Our contributions are largely in three-fold:

1) We propose a novel end-to-end learning framework for the domain adaptation and HPE simultaneously. The domain adaptation network is trained under weak supervision by 2D object segmentation masks and 3D pose labels for hand-only data. Without using 3D annotated HOI data, the method generalizes well and improves the accuracy of 3D hand pose estimation under HOI.

2) The domain adaptation is achieved by two guidances in image space (though they can also be done in feature space). Two image generation methods are investigated and combined: a generative adversarial network and mesh renderer using estimated 3D meshes and *textures*. As an outcome, input HOI images are transformed to segmented and de-occluded hand-only images, effectively improving HPE accuracies.

3) The use of various losses, and the proposed architecture optimises the performance. In addition to the main pipeline, we also investigate the use of real HOI data with its 3D pose labels when available. In extensive experiments for both hand-only and hand-object-interaction, the method outperforms or is on par with state-of-the-arts.

The code is available at the project page[1].

2. Related work

HPE for single hand-only. 3D pose estimation of isolated hands (either from depth images [83, 38, 66, 75, 38, 81, 34, 50, 81, 1, 76] or from RGB images [22, 7, 86, 24]) has achieved great success. Depth-based 3D pose estimation has been well established as depth maps inherently contain 3D information [81, 68, 31], and automatic data synthesis methods [83, 12, 4] help generate large-scale 3D hand pose datasets. In the RGB domain, automatic data generation is much more challenging and it has only recently been tackled using multi-view information and/or differentiable 2D projections [7, 22]: Simon et al. [59] proposed an automatic data annotation scheme which enforces label consistency in a multiple camera setup [24] and Kocabas et al. reconstructed 3D human body skeletons using multi-view 2D skeletons [28]. Further, differentiable renderers and perspective models [6, 15, 21, 31] have enabled training CNNs for 3D mesh reconstruction from single RGB images. They typically employ 2D/3D skeletons and 2D segmentation masks as weak-supervision.

3D hand pose estimation methods can also be categorized into generative and discriminative approaches: Generative methods fit a 3D mesh model to point clouds [72, 69, 54, 67, 65, 64, 51, 42] or intermediate data representations such as 2D skeletons [46]. Most generative methods optimize non-linear data fitting criteria, and therefore susceptible to local optima. With the advent of CNNs and large-scale datasets [59, 81, 68, 81], discriminative methods have shown promising performances and have been established as a strong alternative to generative approaches. However, these methods are agnostic to kinematic and/or geometric (mesh) constraints. Hybrid methods [70, 47, 73, 56] attempt to combine the merits of both discriminative and generative methods. A common strategy in this context is to construct the initialization using discriminative methods and subsequently refine it using generative methods. For example, Tompson et al. [73] apply CNNs to predict hand joint positions and apply particle swarm optimization (PSO) to refine them. Similarly, Sharp et al. [56] estimate the initial joint angles and refine them via PSO. Further exploiting multi-view inputs from strongly interacting hands, Taylor et al. [70] realized a real-time hybrid system.

HPE under hand object interaction (HOI) scenarios. Early works in this domain focused on fitting 3D generative models to RGBD images [45, 18, 30, 19], whereas some works took discriminative approaches, e.g. based on random forests [52, 53]. Model-based hand trackers often suffer from model drift, limiting the range of applicable HOI scenarios. Multi-camera systems have also been exploited [43, 44].

Recently, CNNs have been applied to recovering the HOI hand poses from single RGB images [62, 12, 20, 21, 11, 37, 35, 55, 57, 43, 71]. As annotating 3D joints under occlusion is challenging, exploiting synthetic data has been recently investigated (e.g. GANerated [35], SynthHands [37], and *Obman* [21]). However, existing datasets exhibit a high level of artifacts including unrealistic hand poses (when interacting with objects), and the rendered images therein show a considerable gap from real-world images. Real datasets in the HOI scenarios have also been collected for 3D pose estimation [12, 20, 87] and

[1] https://github.com/bsrvision/weak_da_hands

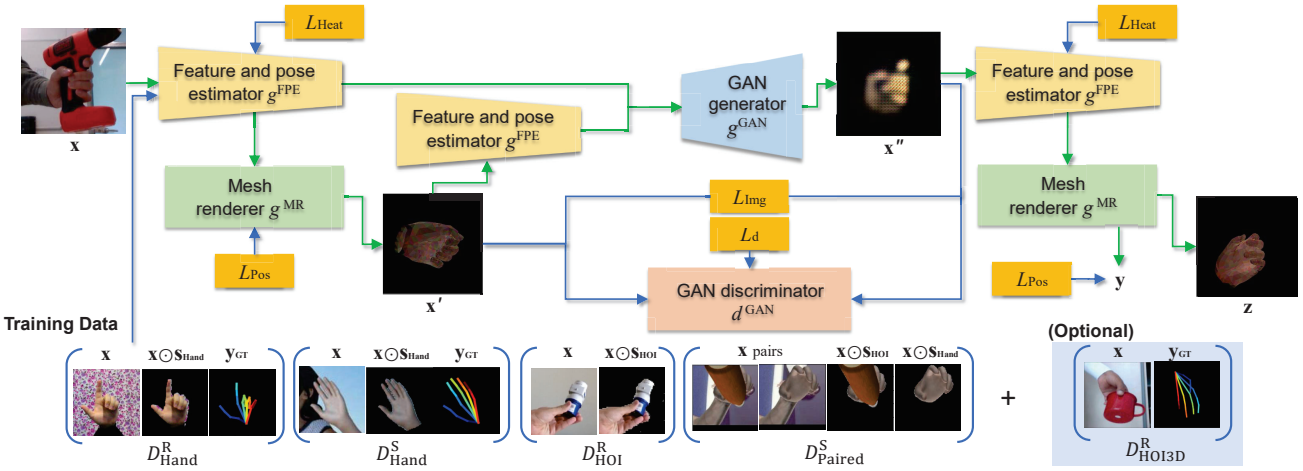


Figure 2: Schematic diagram of the proposed 3D hand mesh and pose estimation framework via domain adaptation. Our domain adaptation network receives an input HOI RGB image $\mathbf{x}$ and extracts 2D feature maps $\mathbf{f}$ and joint heatmaps $\mathbf{h}$ (via 2D feature and pose estimator g^{FPE}). Based on them, the mesh renderer g^{MR} reconstructs the corresponding 3D meshes $\mathbf{m}$ and textures $\mathbf{t}$, and thereafter render these to an initial hand-only image estimate $\mathbf{x}'$. The 2D maps $\{\mathbf{f},\mathbf{h}\}$ and $\{\mathbf{f}',\mathbf{h}'\}$ extracted respectively from $\mathbf{x}$ and $\mathbf{x}'$ are then fed to the GAN generator g^{GAN} which synthesizes a refined hand-only image $\mathbf{x}''$. Finally, g^{FPE} and g^{MR} are applied to $\mathbf{x}''$ to generate the hand free mesh $\mathbf{m}''$ which are then 1) rendered to the corresponding hand-only image $\mathbf{z}$ and 2) used to generate skeletal joint poses $\mathbf{y}$. The green and blue arrows represent the flow of data processing and supervision, respectively.

action recognition [57, 3]. However, they lack in quantity.

The state-of-the-art: Oberweger et al. [41] proposed a feedback loop framework that embeds a depth map generator and uses it to iteratively refine the estimated skeletons. Wei et al. [78] developed a part-based human body pose estimation approach that uses global scene context to compensate for occluded joints. This algorithm generates and gradually refines intermediate 2D heatmap responses. Similar ideas have also been exercised in 3D HOI hand pose estimation (e.g. [35]). However, they require constructing large HOI hand pose datasets. Our algorithm builds upon the architecture of Wei et al. [78] and it achieves superior or comparable performance to the state-of-the-arts without using 3D labels for HOI data. Similarly to our approach, Goudie et al.'s algorithm [16] takes a two-stage approach and uses hand segmentation masks from the HOI images. However, unlike ours, this method does not perform de-occlusion (or inpainting) of the occluded parts, and therefore fails when hands are severely occluded.

There also have been works for tackling the interactions between two hands [74, 36].

Domain adaptation for HPE. Several methods have been developed for reducing the gap between real and synthetic hand data (where only isolated hands appear) [50, 58] or between RGB and depth data [49, 82]. However, to the best of our knowledge, none of prior work has tackled the adaptation of HOI and hand-only domains.

3. Our hand domain adaptation framework

Constructing a pose estimator of hands (HPE) that interact with objects (HOI) is a challenging problem: Existing HPEs trained on hand-only datasets struggle due to object occlusions. Also, training a new HPE under the HOI scenario is not straightforward as the annotated real-world HOI datasets are limited. We propose to mitigate this challenge by mapping the input HOI image to the corresponding object free (hand-only) image, leveraging only easily accessible datasets: the input RGB images in hand-only and HOI scenarios, skeleton annotations for hand-only images, and 2D binary segmentation masks for hand-only and HOI images (which can be extracted based on the accompanying depth maps; See Table 2 for the summary of training datasets and data types that we use).

While this requires restoring (or inpainting) occluded hand regions, which does not have a generally agreed solution, we demonstrate that our framework often faithfully restores occluded hands, and by doing so, it can provide significant performance improvements over existing hand pose estimation approaches.

Overview. Our domain adaptation network (DAN) f^{DAN} receives an input 256×256-sized RGB HOI image $\mathbf{x} \in X$ and generates the corresponding hand-only image $\mathbf{x}' \in X$ and 21 3D skeletal joints $\mathbf{y} \in Y$ estimates. Table 1 provides a summary of notations.

Motivated by the success of recent approaches for hand

Table 1: Summary of notations.

$X \subset \mathbb{R}^{256\times256\times3}$	RGB images ($\mathbf{x}$: input; $\mathbf{x}'$: rendered by g^{MR}; $\mathbf{x}''$: synthesized by g^{GAN}; $\mathbf{z}$: final mesh estimate rendered by g^{MR})
$Y \subset \mathbb{R}^{21\times3}$	3D skeletal pose space
$F \subset \mathbb{R}^{128\times32\times32}$	2D feature space
$H \subset \mathbb{R}^{21\times32\times32}$	2D heatmap space
$M \subset \mathbb{R}^{778\times1538}$	3D mesh space: 778 vertices $\times$ 1,538 faces
$T \subset \mathbb{R}^{1538\times3}$	RGB mesh texture (3 $\times$ 1,538 faces)
$g^{\text{FPE}}: X \rightarrow F \times H$	2D feature and pose estimator
$g^{\text{HME}}: F \times H \rightarrow M$	Hand mesh estimator
$g^{\text{Tex}}: F \times H \rightarrow T$	Texture estimator
$g^{\text{NR}}: M \times T \rightarrow X$	Neural renderer [25]
$g^{\text{Reg}}: M \rightarrow Y$	Hand joint regressor [54]
$g^{\text{MR}}: F \times H \rightarrow X \times Y$	Mesh renderer: $g^{\text{MR}} = [g^{\text{NR}} \circ [g^{\text{HME}}, g^{\text{Tex}}], g^{\text{Reg}} \circ g^{\text{HME}}]$
$g^{\text{GAN}}: F \times H \times F \times H \rightarrow X$	GAN generator
$d_1^{\text{GAN}}, d_2^{\text{GAN}}: X \rightarrow \mathbb{R}$	GAN discriminators
$f^{\text{DAN}}: X \rightarrow X \times Y$	Domain adaptation network: $f^{\text{DAN}} = g^{\text{MR}} \circ g^{\text{FPE}}$

pose estimation [2, 10, 32, 6], we guide the training of DAN by decomposing it into components each provided with intermediate-level supervision: f^{DAN} combines 2D feature and pose estimator (FPE) g^{FPE} and mesh renderer (MR) g^{MR}. FPE extracts 2D spatial feature maps $\mathbf{f} \in F$ and generates heatmaps $\mathbf{h} \in H$ representing the estimated locations of 21 skeletal joints in the image pane. MR consists of 1) hand mesh estimator g^{HME}, 2) texture estimator g^{Tex}, 3) neural renderer g^{NR}, and 4) hand joint regressor g^{Reg}. g^{HME} and g^{Tex} both receive the two outputs $\{\mathbf{f}, \mathbf{h}\}$ of g^{FPE} and estimate the corresponding hand-only mesh $\mathbf{m}$ and texture $\mathbf{t}$, respectively, which are then fed to g^{NR} to synthesize a hand-only image $\mathbf{x}'$. Here, we denote the hand-only image estimated based on $\mathbf{x}$ by $\mathbf{x}'$. Thereafter, g^{Reg} calculates the joint locations $\mathbf{y}$ from $\mathbf{m}$. For g^{NR} and g^{Reg}, we use the models obtained from [25] and [54], respectively. They are held fixed throughout the entire training process of DAN. However, as both are differentiable, they facilitate the training of g^{HME} and g^{Tex} end-to-end.

Training DAN is further guided by generative adversarial networks (GANs): The GAN generator g^{GAN} generates a refined version $\mathbf{x}''$ of $\mathbf{x}'$ conditioned on the features extracted from $\mathbf{x}'$ and $\mathbf{x}$, respectively. The corresponding discriminators $d_1^{\text{GAN}}, d_2^{\text{GAN}}$ are trained to distinguish 1) the synthesized hand-only images and real hand-only images and 2) hand-only images and HOI images, respectively. Figures 2 and 1 show an overview of our DAN architecture and the corresponding examples, respectively.

3.1. 2D feature and pose estimator $g^{\text{FPE}}: X \rightarrow F \times H$

This receives a 256 $\times$ 256-sized RGB image and generates 128 32 $\times$ 32-dimensional spatial feature maps $\mathbf{f}$ and 21 32×32-dimensional heatmaps $\mathbf{h}$, each encoding 2D spatial information at 8 times downsampled resolution from $\mathbf{x}$. The effectiveness of generating such intermediate 2D maps to guide the training of hand and human pose estimators has been demonstrated in [32, 6]. We use the convolutional pose machine (CPM) architecture [78] and the weights pre-trained for hand pose estimation [86]: This algorithm iteratively improves the estimated heatmaps $\mathbf{h}$ by exploiting the corresponding feature maps $\mathbf{f}$ as auxiliary information: The number of total iterations are fixed at 3. Details of all network structures are provided in the accompanying supplemental document.

3.2. Mesh renderer $g^{\text{MR}}: F \times H \rightarrow X \times Y$

Given the 2D maps $\mathbf{f}, \mathbf{h}$ extracted from the input HOI image $\mathbf{x}$, g^{MR} synthesizes the corresponding hand-only image $\mathbf{x}'$ and skeletal joints $\mathbf{y}$ based on its component functions g^{HME}, g^{Tex}, g^{NR}, and g^{Reg}: Following [6, 2, 21], we stratify the training of g^{MR} by first estimating a hand mesh $\mathbf{m}$ as proxy geometric features and then render it to a 2D image $\mathbf{x}'$. The MANO hand model is used to facilitate this process [54]: Our hand mesh estimator g^{HME} first estimates a 63-dimensional MANO parameter vector $\mathbf{p}$ and then converts it to a MANO mesh $\mathbf{m}$ using a differentiable MANO layer g^{MANO}. In parallel, the texture estimator g^{Tex} receives $\mathbf{f}, \mathbf{h}$ and calculates the corresponding mesh color values $\mathbf{t}$. Finally, g^{NR} projects $\mathbf{m}$ and $\mathbf{t}$ onto the image pane to generate the hand-only counter part $\mathbf{x}'$ of the input HOI image $\mathbf{x}$, and g^{Reg} determines the 3D skeletal joints $\mathbf{y}$ from $\mathbf{m}$. Details of the operations of g^{HME}, g^{Tex}, g^{Reg}, and g^{NR} are provided in the supplemental document.

3.3. GAN generator $g^{\text{GAN}}: F \times H \times F \times H \rightarrow X$

Our GAN generator g^{GAN} provides an adaptation of images $\mathbf{x}$ in the HOI domain to the corresponding hand-only images $\mathbf{x}''$, and it has the standard encoder/decoder architecture commonly used in unpaired image translation [85, 23]. Inspired by the work of Kossaifi et al. [29] which involves key points as auxiliary geometric information to improve facial image synthesis, the features $\mathbf{f}, \mathbf{h}$ extracted from $\mathbf{x}$ are used as inputs. Further, we augment the input by adding the features $\mathbf{f}', \mathbf{h}'$ extracted from $\mathbf{x}'$: By adopting the MANO hand model, g^{MR} can synthesize clean hand-only images $\mathbf{x}'$ free of occlusions and background clutters. Even though $\mathbf{x}'$ might not be perfectly aligned with the underlying ground-truth, being combined with features $\mathbf{f}, \mathbf{h}$ of $\mathbf{x}$, $\mathbf{f}', \mathbf{h}'$ can help g^{GAN} generate realistic hand images. We assessed in preliminary experiments, a more straightforward setting where g^{GAN} is conditioned only on $\mathbf{f}, \mathbf{h}$ without generating $\mathbf{x}'$. The corresponding results were significantly worse than $\mathbf{x}''$ indicating that directly estimating the hand-only counterpart of an HOI image is challenging, and having an initial hand reconstruction $\mathbf{x}'$ guided by mesh reconstruction $\mathbf{m}$ helps refine the search space of g^{GAN} parameters during training.

3.4. Training

DAN is trained end-to-end based on 1) hand-only input images $\mathbf{x}$, and the corresponding 3D skeletal joints and hand segmentation masks $\mathbf{s}$ and 2) HOI images and the corresponding segmentation masks each covering hands *and*

Table 2: Data types for training DAN. $\mathbf{s}_{\text{Hand}}$ and $\mathbf{s}_{\text{HOI}}$ represent the foreground hand(-only) and HOI segmentation masks, respectively. For real-world datasets, the segmentation masks are calculated using the accompanying depth maps. Note we do not use $\mathbf{s}_{\text{Hand}}$ for $D^{\text{R}}_{\text{HOI}}$ as extracting hand-only regions is challenging.

$D^{\text{R}}_{\text{Hand}}=\{(\mathbf{x},\mathbf{s}_{\text{Hand}},\mathbf{y})\}$	Real hand-only data (*STB*)
$D^{\text{S}}_{\text{Hand}}=\{(\mathbf{x},\mathbf{s}_{\text{Hand}},\mathbf{y})\}$	Synthetic hand-only data (SynthHands, *RHD*)
$D^{\text{R}}_{\text{HOI}}=\{(\mathbf{x},\mathbf{s}_{\text{HOI}})\}$	Real HOI data (*CORe50*)
$D^{\text{S}}_{\text{Paired}}=\{(\mathbf{x},\mathbf{x}_*,\mathbf{s}_{\text{Hand}},\mathbf{s}_{\text{HOI}}\}$	Paired synthetic HOI ($\mathbf{x}$) and hand-only ($\mathbf{x}_*$) images (*Obman*)
$D_{\text{Hand}}=[D^{\text{R}}_{\text{Hand}},D^{\text{S}}_{\text{Hand}}]$	Hand-only data
$D=[D_{\text{Hand}},D^{\text{R}}_{\text{HOI}},D^{\text{S}}_{\text{Paired}}]$	All training data that we use

objects. Our algorithm is weakly supervised in that it does not use the ground-truth 3D skeletal joints or hand-only segmentation masks for HOI images. However, optionally, when the 3D joints annotations are provided for the HOI images, our algorithm can also exploit them. The training process is summarized in Fig. 2. This section details supervisory information provided to each component of DAN.

2D heatmap supervision L_{Heat}. g^{FPE}'s CPM architecture [78] is initialized with pre-trained weights provided by Zimmermann and Brox [86] and is refined based on the ground-truth 2D heatmaps $\mathbf{h}_{\text{GT}}$ of hand-only images which are induced from the corresponding 3D skeletal annotations [86]. The corresponding loss is given below:

$$L_{\text{Heat}}([g^{\text{FPE}}]_H|D_{\text{Hand}})=\|g^{\text{FPE}}(\mathbf{x})-\mathbf{h}_{\text{GT}}\|_2^2, \quad (1)$$

where $\|A\|_2$ is the L^2 norm of $\text{vec}[A]$ with $\text{vec}[A]$ vectorizing the input multi-dimensional array A and $[\mathbf{v}]_H$ extracts the 2D heatmap component from the output $\mathbf{v}$ of g^{FPE}. We use the notation '$(\cdot|D)$' to signify the type of dataset D from which individual data instances are sampled. As the estimated heatmaps are iteratively improved (see Sec. 3.1), L_{Heat} is applied to each step of iteration. Further, since g^{FPE} generates the heatmaps three times in the training process (for $\mathbf{x}$, $\mathbf{x}'$, and $\mathbf{x}''$, respectively; See Sec. 3.3 and Fig. 2), L_{Heat} is accordingly applied multiple times: $\mathbf{x}$ in Eq. 1 is replaced by $\mathbf{x}'$ and $\mathbf{x}''$, respectively.

Image-level supervision L_{Img}. Each input image $\mathbf{x}$ in the training dataset D is provided with the corresponding 2D segmentation mask $\mathbf{s}$ enabling to extract foreground hand regions $\mathbf{x}\odot\mathbf{s}$ with $\odot$ being element-wise product. For hand-only images, we penalize the deviation between these foreground hands and the corresponding hand-only reconstructions $\mathbf{x}'$ and $\mathbf{x}''$ generated by g^{MR} and g^{GAN}, respectively. Additional supervison is provided via two GAN discriminators: d^{GAN}_1 and d^{GAN}_2 take images $\mathbf{x},\mathbf{x}',\mathbf{x}''$ and respectively distinguish between 1) real and synthesized images and 2) HOI and hand-only images: Images in the hand-only datasets (*STB*, *RHD*, *SH*: see Sec. 4) and HOI datasets (*CORe50*) are used as real images for the first and the second tasks, respectively, while the images generated by g^{GAN} are used as fake images in both tasks. This image-level supervision information is encoded in the loss L_{Img}:

$$\begin{aligned}&L_{\text{Img}}(g^{\text{FPE}},g^{\text{HME}},g^{\text{Tex}}|D)\\&=\sum_{i=1}^{2}\mathbb{E}[\log(1-d_i^{\text{GAN}}(\mathbf{x}''))]+\mathbb{E}[\log(1-d_i^{\text{GAN}}(\mathbf{x}'))]\\&+\|\mathbf{x}''-\mathbf{x}\odot\mathbf{s}_{\text{Hand}}\|_1+\|\mathbf{x}'-\mathbf{x}\odot\mathbf{s}_{\text{Hand}}\|_1. \quad (2)\end{aligned}$$

The last two terms are not used for real HOI images ($D^{\text{R}}_{\text{HOI}}$: see Table 2) as they do not have the corresponding hand-only segmentation masks $\mathbf{s}_{\text{Hand}}$. The discriminator d^{GAN} is further supervised using an adversarial loss:

$$\begin{aligned}&L_d(d^{\text{GAN}}|D)=-\mathbb{E}[\log([d_2^{\text{GAN}}(\mathbf{x}\odot\mathbf{s}_{\text{Hand}})])]\\&-\mathbb{E}[\log(1-[d_2^{\text{GAN}}(\mathbf{x}\odot\mathbf{s}_{\text{HOI}})])] \quad (3)\\&-\mathbb{E}[\log([d_1^{\text{GAN}}(\mathbf{x}\odot\mathbf{s}_{\text{Hand}})])]-\mathbb{E}[\log(1-[d_1^{\text{GAN}}(\mathbf{x}'')])],\end{aligned}$$

where the masked images $\mathbf{x}\odot\mathbf{s}$ are used whenever available, e.g. $D^{\text{S}}_{\text{HOI}}$ provides both $\mathbf{s}_{\text{Hand}}$ and $\mathbf{s}_{\text{HOI}}$ while for $D^{\text{R}}_{\text{Hand}}$, $D^{\text{S}}_{\text{Hand}}$ and $D^{\text{R}}_{\text{HOI}}$, either $\mathbf{s}_{\text{Hand}}$ or $\mathbf{s}_{\text{HOI}}$ are provided.

3D skeleton supervision L_{Pos}. For hand-only data, ground-truth 3D skeletons $\mathbf{y}_{\text{GT}}$ are provided to g^{FPE} and g^{HME}:

$$L_{\text{Pos}}(g^{\text{FPE}},g^{\text{HME}}|D_{\text{Hand}})=\|[g^{\text{MR}}(g^{\text{FPE}}(\mathbf{x}))]_Y-\mathbf{y}_{\text{GT}}\|_2^2, \quad (4)$$

where $[\mathbf{v}]_Y$ extracts the skeleton components from the output $\mathbf{v}$ of g^{MR}. Similarly to L_{Heat}, L_{Pos} is applied twice in each pass of the training images (for $\mathbf{x}$ and $\mathbf{x}'$; the latter case replaces $\mathbf{x}$ in Eq. 4 by $\mathbf{x}'$).

Training sequence. We observed in our preliminary experiments that initializing g^{FPE} and g^{HME} helps the convergence of the training process. Therefore, we train them for the first 30 epochs using L_{Heat} and L_{Pos}. Thereafter, all component functions are jointly trained based on the combined loss L (see our supplemental document for the details of the training process):

$$L=L_{\text{Heat}}+L_{\text{Pos}}+L_{\text{Img}}+L_d. \quad (5)$$

3.5. Testing

Applying the trained DAN to a test image $\mathbf{x}$ follows the same pipeline as training (Fig. 2) except that at testing, no supervision is provided. Once estimated, the MANO parameter vector $\mathbf{p}$ of $\mathbf{x}$ uniquely determines the corresponding hand mesh $\mathbf{m}$ and skeletal joints $\mathbf{y}$. In general, these outputs do not have to perfectly agree with intermediate variables that are generated during the estimation of $\mathbf{p}$, and enforcing consistency with them can help improve $\mathbf{p}$ and thereby generate more accurate mesh and skeleton estimates (see [2] for

Table 3: Error rates of different algorithms on *HO3D* (lower is better) used in Task 3 of HANDS 2019 challenge [8]. The results of all algorithms that participated in this challenge are displayed. The three best results are highlighted with **boldface blue**, *italic green*, and plain orange fonts, respectively.

Participant ID	EXTRAP	OBJECT
potato	*24.74*	27.36
Nplwe	29.19	**18.39**
lin84	31.51	30.59
yhasson	38.42	31.82
LSL	41.81	72.70
SehoKim	49.64	53.79
myunggi	57.45	54.81
sirius.xie	80.06	45.34
Ours	28.24	25.93
Ours + *HO3D* 3D annot.	**23.63**	*20.59*

a similar idea exercised in a different context). In particular, checking and enforcing the consistency of $\mathbf{p}$ and $\mathbf{x}''$ is straightforward as g^{FPE}, g^{HME}, and g^{Reg} are all differentiable with respect to $\mathbf{p}$: Suppose that $\mathbf{j}$ is the 2D skeletal joints recovered by 2D heatmaps $\mathbf{h}''$ estimated on $\mathbf{x}''$. Then, we iteratively refine $\mathbf{p}(0) := \mathbf{p}$ based on the following update rule:

$$\mathbf{p}(t+1) = \mathbf{p}(t) - \gamma \cdot \nabla_{\mathbf{p}} \left(\| [\mathbf{y}]_{XY} - \mathbf{j} \|_2^2 \right), \tag{6}$$

where $[\mathbf{y}]_{XY}$ represents the projection of 3D joints $\mathbf{y}$ onto the image pane (note that $[\mathbf{y}]_{XY}$ depends on $\mathbf{p}(t)$). The number of iterations T and γ are fixed at 50 and 0.01, respectively. This corresponds to the first 50 iterations of the steepest descent on the energy $E(\mathbf{p}) = \| [\mathbf{y}]_{XY}(\mathbf{p}) - \mathbf{j} \|_2^2$ using $\mathbf{p}(0)$ and γ as the initialization and the corresponding step size respectively. The final solution $\mathbf{p}(T)$ improves the average hand pose estimation accuracy from $\mathbf{p}(0)$ by 4.3% and 6.5% on *DO* and *ED* datasets, respectively: 'Ours' and 'Ours (wo/ test refine.)' in Fig. 4(a-b) represent the results obtained from $\mathbf{p}(T)$ and $\mathbf{p}(0)$, respectively.

4. Experiments

We evaluated our algorithm on three datasets in 1) the HOI scenario where weak supervision is provided. Also, we performed experiments on 2) hand-only images, to confirm that our method realizes comparable performance to existing approaches that are designed for the hand-only scenario and 2) on HOI images, after we retrain our system on fully annotated HOI 3D skeletons. The latter set of experiments demonstrate that once available, our algorithm is capable of fully exploiting 3D skeletons and thereby significantly outperforming existing approaches targeted at this setting.

Testing data. We use three challenging real-world HOI datasets: Dexter-object (*DO*) [62], Ego-Dexter (*ED*) [37], and Hand Object-3D (*HO3D*) [20]. Further, to evaluate the performance in the hand-only scenarios, we use the testing split of *STB*. *DO* provides 3,145 video frames sampled from 6 video sequences recording a person interacting with an object and the corresponding 3D fingertip annotations (see Fig. 5(*DO*) for examples). *ED* contains 3,190 RGBD images of hands interacting with 6 objects captured at ego-centric viewpoints. For a subset of 1,485 images in this dataset, 3D fingertip positions are annotated, and we use these labeled images for testing. *STB* has 15,000 hand-only testing frames of a single subject and the corresponding 21 3D skeleton joint annotations. *HO3D* has 6,636 video frames collected in the HOI scenario providing examples of severe object occlusions. Each frame is provided with 21 3D skeleton joint annotations. *HO3D* was originally used in the Task 3 of the HANDS 2019 Challenge [8] and it provides settings for multiple experimental scenarios. Among them, we use 'OBJECT' and 'EXTRAP' configurations which target at assessing the performance of hand pose estimators under varying object categories and hand pose/shape/object category combinations, respectively.

Training data. Our training set consists of real hand-only data $D^{\text{R}}_{\text{Hand}}$, synthetic hand-only data $D^{\text{S}}_{\text{Hand}}$, real HOI data $D^{\text{R}}_{\text{HOI}}$, and synthetic pairs of HOI and hand-only images $D^{\text{S}}_{\text{paired}}$. All images in these training subsets are provided with the corresponding foreground segmentation masks (hand masks for hand-only images and hand+object masks for HOI images). The images in $D^{\text{R}}_{\text{Hand}}$ and $D^{\text{S}}_{\text{Hand}}$ are also accompanied by skeletal joint annotations (see Table 2).

To facilitate direct comparisons with existing work, we used different training set combinations per test dataset: To test on *DO* and *STB*, the training splits of *STB* and *RHD* are used as the hand-only training set D_{Hand}, while for *ED* test set, *STB*, *RHD*, and *SH* are used. Note that *SH* contains both hand-only and HOI images, each provided with ground-truth 3D skeletons. For training, we use only hand-only images of *SH*, while Iqbal et al.'s algorithm [22] that we compare with, uses all images and 3D skeleton annotations. For *HO3D* test set, *STB*, *RHD* and *SH* are used as D_{Hand}.

Evaluation methods. We evaluated our algorithm based on the ratio of correct keypoints (PCK) with varying thresholds and area under the curve (AUC) (of the PCK curves), and compared it with 4 state-of-the-art RGB-based 3D hand pose estimation algorithms [86, 35, 22, 1]: Mueller et al.'s algorithm [35] is tailored for the HOI scenario basing on synthetic images and the corresponding 3D joint annotations in their *GANerated* dataset [35]. Iqbal et al.'s algorithm [22] builds upon the heat map-based framework of [78] and uses a latent depth map generation module that helps recover 3D maps from 2D heat map responses.

On *STB* consisting of hand-only data, we compared with 7 state-of-the-art algorithms designed for the hand-only domain [61, 86, 46, 22, 26, 47, 63] in addition to [22].[2] For *ED*, apart from [22], we are not aware of any existing

[2]The accompanying supplemental document provides a detailed discussion of the algorithms that are compared.

Figure 3: Example hand only image restoration results: (left) input HOI images, (middle) and (right) restored hand images generated by GANs trained with only d^{GAN} (middle) and with d^{GAN}, g^{FPE}, g^{MR} (right), respectively.

work that assesses the accuracies in 3D PCK. To facilitate comparisons with existing work, we also measured the accuracies of our method in 2D PCK.

Results. Figure 4 summarizes the results. On *DO* (Fig. 4a), our algorithm significantly outperformed Mueller et al.'s algorithm [35] and Iqbal et al.'s algorithm [22] both of which were trained on the 3D HOI skeleton annotations.

Interestingly, Baek et al.'s hand mesh reconstruction algorithm [2] also showed significant improvements over [35, 22] even though the former was not designed for the HOI scenarios. This can be attributed to the fact that by employing an explicit 3D hand shape model (the MANO model similarly to ours), [2] provides robustness against moderate occlusions, and further, most input images in *DO* exhibit such mild object occlusions. When the error threshold is larger than $40mm$ (which is typically the case when hands show severe occlusions), our algorithm consistently outperformed [2], and it becomes comparable to [2] as the threshold decreases.

Figure 4(b-c) show the results for the *ED* dataset: When evaluated in 3D PCK, our algorithm clearly outperformed Iqbal et al.'s algorithm. In 2D PCK, the performance of our algorithm is comparable to [22, 35] and it outperforms [86] which uses hand-only images.

On *STB*, ours is again significantly better than or comparable to state-of-the-art algorithms [61, 86, 46, 22, 26, 47, 63] demonstrating that it continues to offer state-of-the-art performance even in the hand-only domain (Fig. 4(d)). Note that directly applying the system trained on the HOI data (Mueller et al.'s algorithm [35]) to hand-only images can significantly degrade the performance. Iqbal et al.'s algorithm [22] is retrained on hand-only images.

Table 3 shows the results of our algorithm compared with (all) eight algorithms that participated in the Task 3 of HANDS 2019 challenge which used *HO3D*. Our algorithm achieved the best and second-best results for EXTRAP and OBJECT, respectively (disregarding Ours + *HO3D* 3D annot.) further confirming our initial thesis that it can provide stare-of-the-art performance in the HOI scenarios even without requiring 3D skeletal annotations and/or known object types: It should be noted that our algorithm has been trained on datasets that are disjoint from *HO3D* and there is no overlap between the object categories of *HO3D* and our training datasets. Further, when provided with such skeletal annotations, our algorithm ranked the best and second best for EXTRAP and OBJECT, respectively demonstrating its ability to fully accommodate such high-quality supervision.

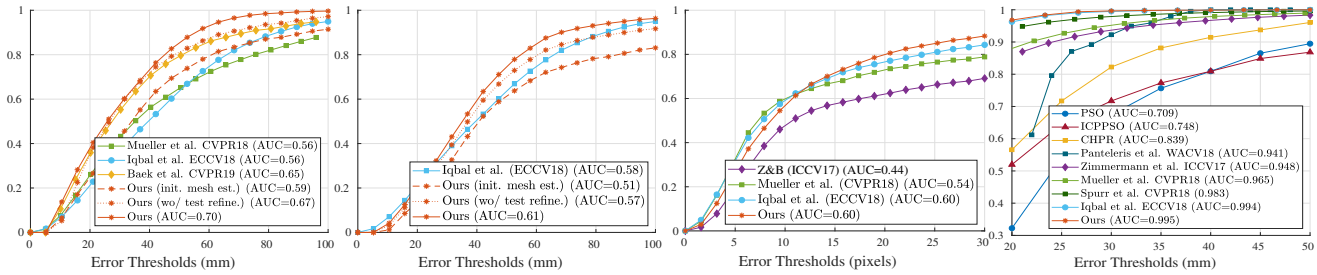
Figure 5 exemplifies how such a significant performance gain can be achieved using only weak-supervision: By employing the MANO 3D hand model and GAN generators, and iteratively enforcing the consistency of the final mesh reconstruction over 2D maps, our algorithm can faithfully recover the hand-only counterpart of the input HOI image. It should be noted that the initial errors in hand-only restoration $\mathbf{x}'$ has been subsequently corrected via GAN generators $\mathbf{x}''$ and/or mesh refinement performed at the testing ($\mathbf{z}$).

Ablation study. To gain an insight into the contribution of our system components, we measured the pose estimation accuracies of our system that uses 1) only the initial mesh reconstruction $\mathbf{x}'$ (without the subsequent explicit domain adaptation steps) and 2) the final mesh estimation without the testing refinement step. The corresponding results ('Ours (init. mesh est.)' and 'Ours (wo/ test refine.)' in Fig. 4 (a, b) for *DO* and *ED* datasets, respectively) show that both domain adaptation steps, and the iterative refinement step contribute significantly in improving the performance.

We also trained a separate instance of our system on the training splits of *HO3D* using the corresponding 3D annotations to assess its performance when provided with 3D skeletons. In Table 3, the resulting system is denoted as 'Ours+*HO3D* 3D annot.' while our final system trained without *HO3D*'s 3D skeletal annotations is denoted as 'Ours'.

Discussion. The hand-only images synthesized by our GANs are rather blurry. This can be attributed to the fact that in our framework, the GAN parameters are updated based not only on the discriminator d^{GAN} but also on the 2D feature and pose estimator g^{FPE} and mesh renderer g^{MR}: The latter two components do not promote sharpness in the final results. When we remove these two components, corresponding reconstructions are not blurry but the final hand pose estimates become less accurate. The corresponding synthesized examples are shown in Fig. 3.

Our PyTorch-based implementation takes 600ms and 2s per frame on a single NVIDIA Geforce GTX1080 Ti GPU and on a Geforce GTX 1050 mobile GPU, respectively.



(a) 3D PCK on *DO* (b) 3D PCK on *ED* (c) 2D PCK on *ED* (d) 3D PCK on *STB*

Figure 4: Performances of different algorithms on three benchmark datasets: (a) *DO*, (b-c) *ED*, (d) *STB*, respectively.

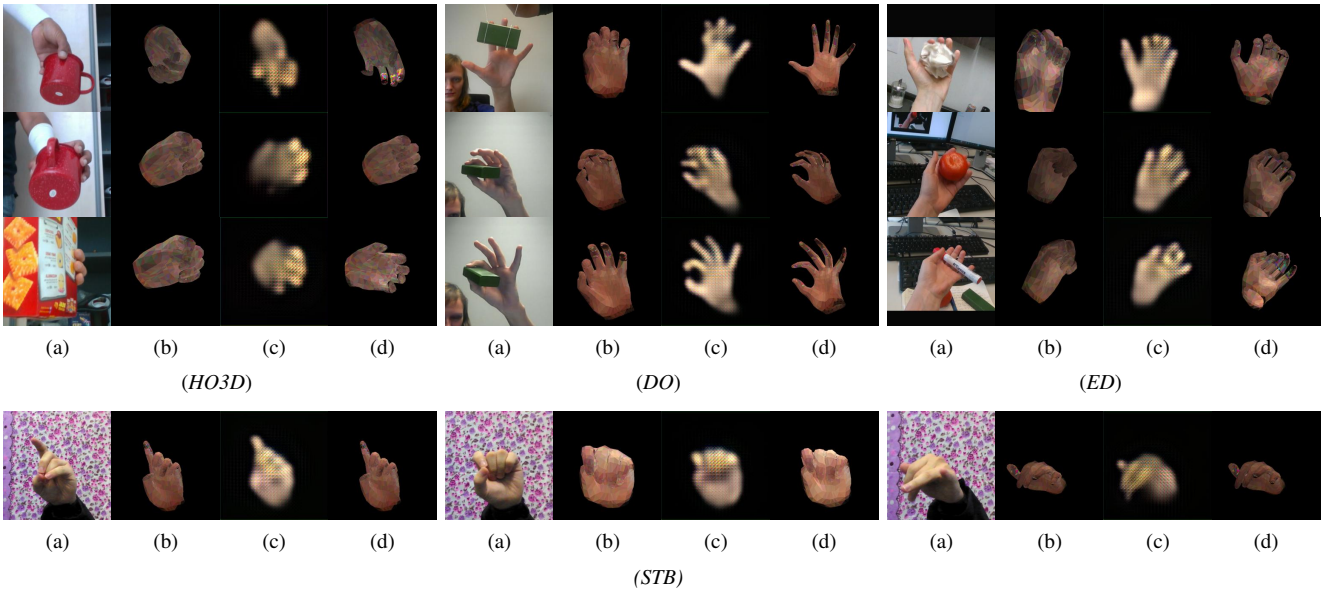


Figure 5: Example hand image restoration results via domain adaptation under the HOI (*HO3D*, *DO*, *ED*) and hand-only (*STB*) scenarios. (a) input images $\mathbf{x}$, (b) images $\mathbf{x}'$ generated by our initial mesh renderer g^{MR}, (c) images $\mathbf{x}''$ generated by our GAN generator g^{GAN}, (d) final images $\mathbf{z}$ generated by the mesh renderer g^{MR}.

5. Conclusions

Existing approaches to estimating skeletons of hands that interact with objects require fully annotated 3D skeletal joints, which are costly to build due to object occlusions. We have presented a new framework that trains an estimator without having to construct such fully annotated datasets. The crux of our approach is a new domain adaptation framework that transfers input HOI images to the corresponding hand-only images only based on 2D foreground segmentation masks, 3D skeletons for hand-only images, and synthetic hand-only and HOI image pairs, all of which can be easily constructed for synthetic datasets or real RGB images accompanied by depth maps. We designed a new training process that fully leverages such weak supervision in an end-to-end manner. Evaluated on 3 real-world HOI datasets and a hand-only dataset, we demonstrated that 1) on HOI images, our algorithm provides superior or comparable performance to existing approaches that are trained on fully annotated skeletons and 2) ours still retains state-of-the-art performance on hand-only images. Further, 3) when provided with optional skeleton annotations, it can significantly outperform existing HOI pose estimation approaches.

As our method is learning-based, it suffers from a new-type of testing data. We illustrated the "yellow bottle" as one such data in the supplemental. Furthermore, ours does not exploit 1) temporal information and 2) the class/shape of objects under interaction. Future work should explore the possibilities of enforcing temporal consistency over the estimated mesh shapes and refining pose estimates via exploiting (recognized) action contexts exercised on objects under interaction.

Acknowledgement. Kwang In Kim was supported by National Research Foundation of Korea (NRF) grant funded by Korea government (MSIT): NRF-2019R1F1A1061603.

References

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Augmented skeleton space transfer for depth-based hand pose estimation. In *CVPR*, 2018. 1, 2, 6

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In *CVPR*, 2019. 4, 5, 7

[3] Seungryul Baek, Zhiyuan Shi, Masato Kawade, and Tae-Kyun Kim. Kinematic-layout-aware random forests for depth-based action recognition. In *BMVC*, 2017. 3

[4] Binod Bhattarai, Seungryul Baek, Rumeysa Bodur, and Tae-Kyun Kim. Sampling strategies for GAN synthetic data. In *ICASSP*, 2020. 2

[5] Abhishake Kumar Bojja, Franziska Mueller, Sri Raghu Malireddi, Markus Oberweger, Vincent Lepetit, Christian Theobalt, Kwang Moo Yi, and Andrea Tagliasacchi. HandSeg: An automatically labeled dataset for hand segmentation from depth images. In *Conference on Computer and Robot Vision (CRV)*, 2019. 1

[6] Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 3D hand shape and pose from images in the wild. In *CVPR*, 2019. 2, 4

[7] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, 2018. 1, 2

[8] ICCV HANDS2019 Challenge. https://sites.google.com/view/hands2019/challenge, (accessed 15 Nov. 2019). 6

[9] Tzu-Yang Chen, Pai-Wen Ting, Min-Yu Wu, and Li-Chen Fu. Learning a deep network with spherical part model for 3D hand pose estimation. In *ICRA*, 2017. 1

[10] Yujin Chen, Zhigang Tu, Liuhao Ge, Dejun Zhang, Ruizhi Chen, and Junsong Yuan. SO-HandNet: Self-organizing network for 3D hand pose estimation with semi-supervised learning. In *ICCV*, 2019. 4

[11] Chiho Choi, Sang Ho Yoon, Chin-Ning Chen, and Karthik Ramani. Robust hand pose estimation during the interaction with an unknown object. In *ICCV*, 2017. 1, 2

[12] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *CVPR*, 2018. 1, 2

[13] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3D Hand pose estimation in single depth images: from single-view CNN to multi-view CNNs. In *CVPR*, 2016. 1

[14] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *CVPR*, 2017. 1

[15] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3D hand shape and pose estimation from a single RGB image. In *CVPR*, 2019. 2

[16] Duncan Goudie and Aphrodite Galata. 3D hand-object pose estimation from depth with convolutional neural networks. In *FG*, 2017. 3

[17] Hengkai Guo, Guijin Wang, Xinghao Chen, Cairong Zhang, Fei Qiao, and Huazhong Yang. Region ensemble network: improving convolutional network for hand pose estimation. In *ICIP*, 2017. 1

[18] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *CVPR*, 2010. 2

[19] Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. Tracking a hand manipulating an object. In *ICCV*, 2009. 2

[20] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. HO-3D: A multi-user, multi-object dataset for joint 3D hand-object pose estimation. In *ArXiv*, 2019. 1, 2, 6

[21] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 3D hand shape and pose estimation from a single RGB image. In *CVPR*, 2019. 1, 2, 4

[22] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5D heatmap regression. In *ECCV*, 2018. 1, 2, 6, 7

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image translation with conditional adversarial networks. In *CVPR*, 2017. 4

[24] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: a massively multiview system for social motion capture. In *ICCV*, 2015. 2

[25] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3D mesh renderer. In *CVPR*, 2018. 4

[26] James Kennedy and Russell Eberhart. Particle Swarm Optimization. In *IEEE International Conference on Neural Networks*, 1995. 6, 7

[27] Cem Keskin, Furkan Kıraç, Yunus Emre Kara, and Lale Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *ECCV*, 2012. 1

[28] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3D human pose using multi-view geometry. In *CVPR*, 2019. 2

[29] Jean Kossaifi, Linh Tran, Yannis Panagakis, and Maja Pantic. GAGAN: Geometry-aware generative adversarial networks. In *CVPR*, 2018. 4

[30] Nikolaos Kyriazis and Antonis A. Argyros. Scalable 3D tracking of multiple interacting objects. In *CVPR*, 2014. 2

[31] Jameel Malik, Ahmed Elhayek, Fabrizio Nunnari, Kiran Varanasi, Kiarash Tamaddon, Alexis Heloir, and Didier Stricker. DeepHPS: End-to-end estimation of 3D hand pose and shape by learning from synthetic depth. In *3DV*, 2018. 2

[32] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 4

[33] Stan Melax, Leonid Keselman, and Sterling Orsten. Dynamics based 3D skeletal hand tracking. In *i3D*, 2013. 1

[34] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. In *CVPR*, 2018. 1, 2

[35] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian

Theobalt. GANerated hands for real-time 3D hand tracking from monocular RGB. In *CVPR*, 2018. 1, 2, 3, 6, 7

[36] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *SIGGRAPH*, 2019. 3

[37] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *ICCV*, 2017. 1, 2, 6

[38] Markus Oberweger and Vincent Lepetit. DeepPrior++: Improving fast and accurate 3D hand pose estimation. In *ICCV HANDS Workshop*, 2017. 1, 2

[39] Markus Oberweger, Gernot Riegler, Paul Wohlhart, and Vincent Lepetit. Efficiently creating 3D training data for fine hand pose estimation. In *CVPR*, 2016. 1

[40] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. In *Proc. Computer Vision Winter Workshop*, 2015. 1

[41] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Generalized feedback loop for joint hand-object pose estimation. *TPMAI*, 2019. 3

[42] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Efficient model-based 3D tracking of hand articulations using kinect. In *BMVC*, 2011. 1, 2

[43] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, 2011. 2

[44] Paschalis Panteleris and Antonis Argyros. Back to RGB: 3D tracking of hands and hand-object interactions based on short-baseline stereo. In *ICCV HANDS Workshop*, 2017. 2

[45] Paschalis Panteleris, Nikolaos Kyriazis, and Antonis A. Argyros. 3D tracking of human hands in interaction with unknown objects. In *BMVC*, 2015. 2

[46] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single RGB frame for real time 3D hand pose estimation in the wild. In *WACV*, 2018. 2, 6, 7

[47] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *CVPR*, 2014. 2, 6, 7

[48] Kha Gia Quach, Chi Nhan Duong, Khoa Luu, and Tien D. Bui. Depth-based 3D hand pose tracking. In *ICPR*, 2016. 1

[49] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Domain transfer for 3D pose estimation from color images without manual annotations. In *ACCV*, 2018. 3

[50] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Feature mapping for learning fast and accurate 3D pose inference from synthetic images. In *CVPR*, 2018. 1, 2, 3

[51] Konstantinos Roditakis, Alexandros Makris, and Antonis A. Argyros. Generative 3D hand tracking with spatially constrained pose sampling. In *BMVC*, 2017. 1, 2

[52] Gregory Rogez, James S. Supancic, and Deva Ramanan. First-person pose recognition using egocentric workspaces. In *CVPR*, 2015. 1, 2

[53] Gregory Rogez, James S. Supancic, and Deva Ramanan. Understanding everyday hands in action from RGB-D images. In *ICCV*, 2015. 2

[54] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. In *SIGGRAPH Asia*, 2017. 2, 4

[55] Matthias Schröder and Helge Ritter. Hand-object interaction detection with fully convolutional networks. In *CVPR*, 2017. 2

[56] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, and S. Izadi. Accurate, robust and flexible real-time hand tracking. In *CHI*, 2015. 2

[57] Roy Shilkrot, Supreeth Narasimhaswamy, Saif Vazir, and Minh Hoai Nguyen. Working Hands: a hand-tool assembly dataset for image segmentation and activity mining. In *BMVC*, 2019. 2, 3

[58] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017. 3

[59] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 1, 2

[60] Ayan Sinha, Chiho Choi, and Karthik Ramani. DeepHand: Robust hand pose estimation by completing a matrix imputed with deep features. In *CVPR*, 2016. 1

[61] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 1, 6, 7

[62] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from RGB-D input. In *ECCV*, 2016. 1, 2, 6

[63] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *CVPR*, 2015. 6, 7

[64] Jonathan Talyor, Richard Stebbing, Varun Ramakrishna, Cem Keskin, Jamie Shotton, Shahram Izadi, Aaron Hertzmann, and Andrew Fitzgibbon. User-specific hand modeling from monocular depth sequences. In *CVPR*, 2014. 2

[65] David Joseph Tan, Thomas Cashman, Jonathan Taylor, Andrew Fitzgibbon, Daniel Tarlow, Sameh Khamis, Shahram Izadi, and Jamie Shotton. Fits like a glove: Rapid and reliable hand shape personalization. In *CVPR*, 2016. 1, 2

[66] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: structural estimation of 3D articulated hand posture. *TPAMI*, 2016. 1, 2

[67] Danhang Tang, Jonathan Taylor, Pushmeet Kohli, Cem Keskin, Tae-Kyun Kim, and Jamie Shotton. Opening the black box: hierarchical sampling optimization for estimating human hand pose. In *ICCV*, 2015. 2

[68] Danhang Tang, Tsz-Ho Yu, and Tae-Kyun Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *ICCV*, 2013. 1, 2

[69] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David

Sweeney, Julien Valentin, Benjamin Luff, Arran Topalian, Erroll Wood, Sameh Khamis, Pushmeet Kohli, Shahram Izadi, Richard Banks, Andrew Fitzgibbon, and Jamie Shotton. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. In *SIGGRAPH*, 2016. 1, 2

[70] Jonathan Taylor, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Articulated distance fields for ultra-fast tracking of hands interacting. *ToG*, 2017. 2

[71] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *CVPR*, 2019. 2

[72] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. In *SIGGRAPH Asia*, 2016. 1, 2

[73] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *TOG*, 2014. 2

[74] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016. 3

[75] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Crossing nets: dual generative models with a shared latent space for hand pose estimation. In *CVPR*, 2017. 1, 2

[76] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Self-supervised 3D hand pose estimation through training by fitting. In *CVPR*, 2019. 2

[77] Chengde Wan, Angela Yao, and Luc Van Gool. Hand Pose Estimation from Local Surface Normals. In *ECCV*, 2016. 1

[78] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 3, 4, 5, 6

[79] Min-Yu Wu, Ya Hui Tang, Pai-Wei Ting, and Li-Chen Fu. Hand pose learning: combining deep learning and hierarchical refinement for 3D hand pose estimation. . In *BMVC*, 2017. 1

[80] Qi Ye, Shanxin Yuan, and Tae-Kyun Kim. Spatial attention deep net with partial PSO for hierarchical hybrid hand pose estimation. In *ECCV*, 2016. 1

[81] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, Junsong Yuan, Xinghao Chen, Guijin Wang, Fan Yang, Kai Akiyama, Yang Wu, Qingfu Wan, Meysam Madadi, Sergio Escalera, Shile Li, Dongheui Lee, Iason Oikonomidis, Antonis Argyros, and Tae-Kyun Kim. Depth-based 3D hand pose estimation: From current achievements to future goals. In *CVPR*, 2018. 1, 2

[82] Shanxin Yuan, Björn Stenger, and Tae-Kyun Kim. 3D hand pose estimation from RGB using privileged learning with depth data. In *ICCV Workshop on Hands*, 2019. 3

[83] Shanxin Yuan, Qi Ye, Björn Stenger, Siddhant Jain, and Tae-Kyun Kim. Big hand 2.2M benchmark: hand pose data set and state of the art analysis. In *CVPR*, 2017. 1, 2

[84] Xingyi Zhou, Qingfu Wan, Wei Zhang, Xiangyang Xue, and Yichen Wei. Model-based deep hand pose estimation. In *IJCAI*, 2016. 1

[85] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 4

[86] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, 2017. 1, 2, 4, 5, 6, 7

[87] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHand: A dataset for markerless capture of hand pose and shape from single RGB image. In *ICCV*, 2019. 1, 2