3D Hand Pose Estimation via Regularized Graph Representation Learning

Anonymous Authors
AAAI paper 6661

Abstract

This paper addresses the problem of 3D hand pose estimation from a monocular RGB image. While previous methods have shown great success, the structure of hands has not been fully exploited, which is critical in pose estimation. To this end, we propose a regularized graph representation learning under a conditional adversarial learning framework for 3D hand pose estimation, aiming to capture structural inter-dependencies of hand joints. In particular, we estimate an initial hand pose from a parametric hand model as a prior of hand structure, which regularizes the inference of the structural deformation in the prior pose for accurate graph representation learning via residual graph convolution. To optimize the hand structure further, we propose two bone-constrained loss functions, which characterize the morphable structure of hand poses explicitly. Also, we introduce an adversarial learning framework conditioned on the input image with a multi-source discriminator, which imposes the structural constraints onto the distribution of generated 3D hand poses for anthropomorphically valid hand poses. Extensive experiments demonstrate that our model sets the new state-of-the-art in 3D hand pose estimation from a monocular image on five standard benchmarks.

1 Introduction

3D human hand pose estimation is a long-standing problem in computer vision, which is critical for various applications such as virtual reality and augmented reality (Hürst and van Wezel 2011; Piumsomboon et al. 2013). Previous works attempt to estimate hand pose from depth images (Ge et al. 2016; Wu et al. 2018; Zhou et al. 2018; Ge et al. 2018) or in multi-view setups (Panteleris and Argyros 2017; Zhang et al. 2016a). However, due to the diversity and complexity of hand shape, gesture, occlusion, *etc.*, it still remains a challenging problem despite years of studies (Rehg and Kanade 1994; Ying and Huang 2002; Ying, John, and Huang 2005; Hui et al. 2017).

As RGB cameras are more widely accessible than depth sensors, recent works focus mostly on 3D hand pose estimation from a monocular RGB image and have shown their efficiency (Ge et al. 2019; Boukhayma, Bem, and Torr 2019; Baek, Kim, and Kim 2019; Cai et al. 2018; Zimmermann and Brox 2017a; Doosti et al. 2020). While some early works (Cai et al. 2018; Boukhayma, Bem, and Torr 2019) did not explicitly exploit the structure of hands, some recent methods (Ge et al. 2019; Doosti et al. 2020) have shown the crucial role of hand structure in pose estimation, but may resort to an additional synthetic dataset. Also, unlike bodies and faces that have obvious local characteristics (*e.g.*, eyes on a face), hands exhibit almost uniform appearance. Consequently, estimated hand poses from existing methods are sometimes distorted and unnatural.

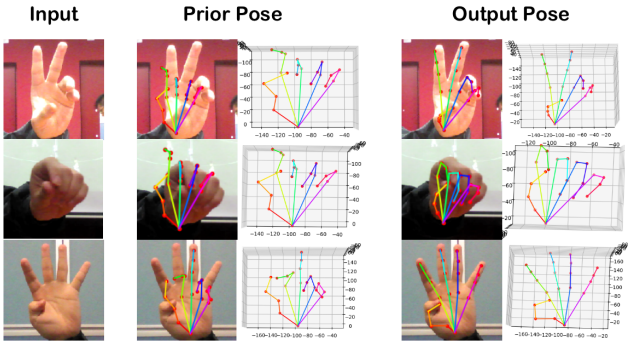

Figure 1: **The proposed method estimates 3D hand pose from a monocular image based on regularized graph representation learning.** A parametric hand model generates a *prior pose*, which regularizes the learning of deformations in graph topology under a conditional adversarial learning framework.

To fully exploit the structure of hands, we propose to represent the irregular topology of 3D hand poses naturally on graphs, and learn the graph representation regularized by a prior pose from the monocular image input under a conditional generative adversarial learning framework, aiming to capture the structural dependencies among hand joints. Based on the Maximum a Posteriori estimation formulation of inferring 3D hand pose, we first construct an initial hand pose from a parametric hand model as a prior of hand structure (*prior pose*), which captures the holistic topology of hand structures, *i.e.*, the adjacency relations between joints. Based on this prior pose, we represent the topology of hand joints on a graph, where each joint is treated as a node and each pair of adjacent nodes are connected. We further learn the *deformation* in the prior pose to refine the hand structure representation, by propagating information across adjacent nodes via residual graph convolution and conditional

on the input image. Moreover, while most existing works (Boukhayma, Bem, and Torr 2019; Ge et al. 2019; Cai et al. 2018) deploy 3D Euclidean distance between joints as the loss function for 3D annotation, we propose two *bone loss functions* that constrain the length and orientation of each bone connected by adjacent joints so as to preserve hand structure explicitly. On the other hand, to address the challenge of uniform appearance, we propose to train the network under an adversarial learning framework conditioned on the input image, aiming to estimate the real distribution of 3D hand poses. Besides, unlike some recent works (Ge et al. 2019; Cai et al. 2018; Kulon et al. 2019), we estimate 3D hand poses *without* resorting to ground truth meshes or depth maps, which is more suitable for datasets in the wild.

Specifically, given an input monocular image, our framework consists of a hand pose generator and a conditional discriminator. The generator is composed of a MANO hand model module (Romero, Tzionas, and Black 2017) that provides an initial pose estimation as prior pose and a deformation learning module regularized by the prior pose. In particular, taking the prior pose and image features as input, the deformation learning module learns the deformation in the prior pose to further refine the hand structure, by our designed residual graph convolution that leverages on the recently proposed ResGCN (Li et al. 2019). Further, we design a conditional multi-source discriminator that employs hand poses, hand bones computed from poses as well as the input image to distinguish the predicted 3D hand pose from the ground-truth, leading to anthropomorphically valid hand pose. Experimental results demonstrate that our model achieves significant improvements over state-of-the-art approaches on five standard benchmarks.

To summarize, our main contributions include

- We propose regularized graph representation learning for 3D hand pose estimation from a monocular image, which fully exploits structural dependencies among hand joints.
- We learn the graph representation of hand poses by inferring structural deformation, which is regularized by an initial hand pose estimation from a parametric hand model.
- We introduce two bone-constrained loss functions, which optimize the estimation of hand structures by explicitly enforcing constrains on the topology of bones.
- We present a conditional adversarial learning framework to impose structural constraints onto the distribution of generated 3D hand poses, which is able to address the challenge of uniform appearance in hands.

2 Related Work

According to the input modalities, previous works on 3D hand pose estimation can be classified into three categories: 1) 3D hand pose estimation from depth images; 2) 3D hand pose estimation from multiple RGB images; 3) 3D hand pose estimation from a monocular RGB image.

2.1 Estimation from Depth Images

Depth images contain rich 3D information for hand pose estimation (Tang, Yu, and Kim 2013), which has shown promising accuracy (Yuan et al. 2018). There is a rich literature on 3D hand pose estimation with depth images as input (Ge et al. 2018, 2016; Fitzgibbon 2015; Choi 2016; De, Fleet, and Paragios 2011; Khamis et al. 2015; Xiao et al. 2015; Malik et al. 2018; Oberweger and Lepetit 2018). Among them, some earlier works such as (De, Fleet, and Paragios 2011; Khamis et al. 2015) are based on a deformable hand model with an iterative optimization training approach. Due to the effectiveness of deep learning, some recent works like (Malik et al. 2018) leverage CNN to learn the shape and pose parameters for a proposed model (LBS hand model).

2.2 Estimation from Multiple Images

Multiple RGB images taken from different views also contain rich 3D information. Therefore, some works take multiple images as input to alleviate the occlusion problem (Campos and Murray 2006; Oikonomidis, Kyriazis, and Argyros 2010; Sridhar et al. 2014). Campos *et al.* (Campos and Murray 2006) propose a regression-based approach for hand pose estimation, where they utilize multi-view images to overcome the occlusion issue. Sridhar *et al.* (Sridhar et al. 2014) contribute a fundamentally extended generative tracking algorithm based on an augmented implicit shape representation with multiple images as input.

2.3 Estimation from a Monocular Image

Compared with the aforementioned two categories, a monocular RGB image is more accessible. Early works (Athitsos and Sclaroff 2003; Rehg and Kanade 2002; Stenger et al. 2006) propose complex model-fitting approaches, which are based on dynamics and multiple hypotheses and depend on restricted requirements. These model-fitting approaches have proposed many hand models, based on assembled geometric primitives (Oikonomidis, Kyriazis, and Argyros 2011) or sphere meshes (Tkach, Pauly, and Tagliasacchi 2016), *etc.* Our work deploys the MANO hand model (Romero, Tzionas, and Black 2017) as our prior, which models both hand shape and pose as well as generates meshes. Nevertheless, these sophisticated approaches suffer from low estimation accuracy.

With the advance of deep learning, many recent works estimate 3D hand pose from a monocular RGB image using neural networks (Ge et al. 2019; Boukhayma, Bem, and Torr 2019; Baek, Kim, and Kim 2019; Cai et al. 2018; Zimmermann and Brox 2017a; Yang and Yao 2019; Kulon et al. 2019). Among them, some recent works (Kulon et al. 2019; Ge et al. 2019) directly reconstruct the 3D hand mesh and then generate the 3D hand pose through a pose regressor. Kulon *et al.* (Kulon et al. 2019) reconstruct the hand pose based on an auto-encoder, which employs an encoder to extract the latent code and feeds the latent code into the decoder to reconstruct hand mesh. Ge *et al.* (Ge et al. 2019) propose to estimate vertices of 3D meshes from GCNs (Kipf and Welling 2017) in order to learn nonlinear variations in hand shape. The latent feature of the input RGB image is extracted via several networks and then fed into a GCN to directly infer the 3D coordinates of mesh vertices. However, since the accuracy of the output hand mesh is criti-

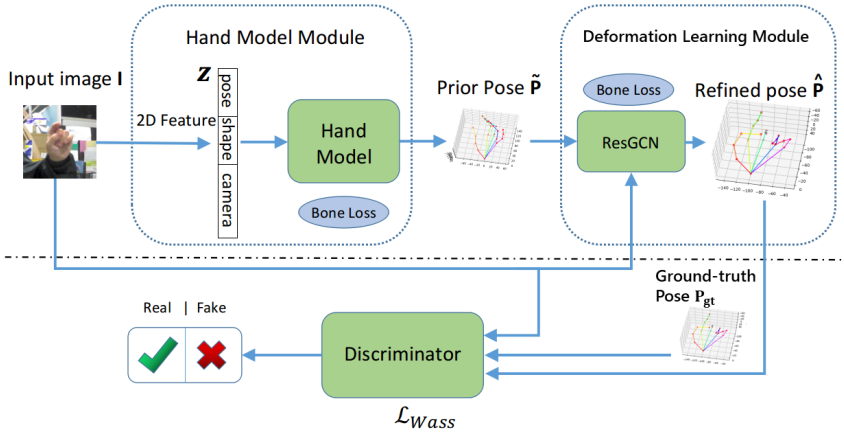

Figure 2: Architecture of the proposed regularized graph representation learning under a conditional adversarial learning framework for 3D hand pose estimation.

cal for both methods, they need an extra dataset which provides ground truth hand meshes as supervision. Also, the upsampling layer used in (Ge et al. 2019) to reconstruct the hand mesh will cause a non-uniform distribution of vertices in mesh, which influences the accuracy of hand pose.

In contrast, we take a prior pose estimated from a parametric hand model as regularization for graph representation learning over hand poses rather than directly reconstructing hand poses from latent features. Besides, our method does not require any additional supervision such as mesh supervision (Ge et al. 2019; Kulon et al. 2019) or depth image supervision (Ge et al. 2019; Cai et al. 2018). Hence, our method is more suitable for datasets in the wild. Further, we introduce conditional adversarial training for 3D hand pose estimation, which enables learning a real distribution of 3D hand poses.

3 Methodology

3.1 Overview of the Proposed Approach

We aim to infer 3D hand pose via regularized graph representation learning under an adversarial learning framework. The entire framework consists of a hand pose generator $\mathbb{G}$ and a conditional discriminator $\mathbb{D}$, as illustrated in Fig. 2.

Given a monocular RGB image $\mathbf{I}$ as the input, the generator $\mathbb{G}$ includes two modules:

- **The hand model module** generates an initial estimation of 3D hand pose $\tilde{\mathbf{P}} \in \mathbb{R}^{N\times 3}$ with N joints ($N = 21$ in our experimental setting), which serves as a *prior pose* for the subsequent refinement. This module consists of a feature extractor and a parametric hand model.
- **The deformation learning module** infers the structural deformation in the prior pose $\tilde{\mathbf{P}}$ for regularized graph representation learning. Taking $\tilde{\mathbf{P}}$ and $\mathbf{I}$ as the input, this module exploits the structural relationship among hand joints via residual graph convolution and outputs the deformation, leading to the refined pose $\hat{\mathbf{P}} \in \mathbb{R}^{N\times 3}$.

The multi-source discriminator $\mathbb{D}$ imposes structural constraints onto the distribution of generated 3D hand poses conditioned on the input image, which distinguishes the ground-truth 3D poses from the predicted ones.

3.2 The Proposed Hand Pose Generator $\mathbb{G}$

Given the observed input image $\mathbf{I}$ and ground truth hand pose $\mathbf{P}_{\text{gt}}$, we formulate the training of hand pose estimation from a monocular image as a Maximum a Posteriori (MAP) estimation problem:

$$\hat{\mathbf{P}}_{\text{MAP}}(\mathbf{I}, \mathbf{P}_{\text{gt}}) = \underset{\mathbf{P}}{\operatorname{argmax}}\, f(\mathbf{I}, \mathbf{P}_{\text{gt}}|\mathbf{P})g(\mathbf{P}), \qquad (1)$$

where $\mathbf{P}$ denotes the hand pose to estimate. In (1), $g(\mathbf{P})$ represents the prior probability distribution of the hand pose, which provides the prior knowledge of $\mathbf{P}$. $f(\mathbf{I}, \mathbf{P}_{\text{gt}}|\mathbf{P})$ denotes the likelihood function, which is the probability of obtaining the observed image $\mathbf{I}$ and ground truth hand pose $\mathbf{P}_{\text{gt}}$ given the estimated hand pose $\mathbf{P}$.

We define the likelihood function as an exponential function of the distance between the estimated pose and the ground truth pose/input image:

$$f(\mathbf{I}, \mathbf{P}_{\text{gt}}|\mathbf{P}) = \exp\{-d_1(\mathbf{P}_{\text{gt}}, \mathbf{P}) - d_2(\mathbf{I}, \mathbf{P})\}, \qquad (2)$$

where $d_1(\cdot)$ is the distance metric between the estimated hand pose and the ground truth, and $d_2(\cdot)$ is the distance metric between the estimated hand pose and the input image. Regarding $g(\mathbf{P})$, it is a constant C after we acquire a prior pose from a parametric hand model. Hence, when we substitute (2) and $g(\mathbf{P}) = C$ into (1), take the logarithm and multiply by -1, we have

$$\min_{\mathbf{P}} d_1(\mathbf{P}_{\text{gt}}, \mathbf{P}) + d_2(\mathbf{I}, \mathbf{P}). \qquad (3)$$

$d_1(\cdot)$ and $d_2(\cdot)$ will be discussed in Section 3.4 in detail.

Specifically, we employ a parametric hand model to provide the prior of $\mathbf{P}$, and designate a Deformation Learning Module to learn the pose under the supervision of the ground-truth pose and input image. We discuss the two modules of the generator in detail as follows.

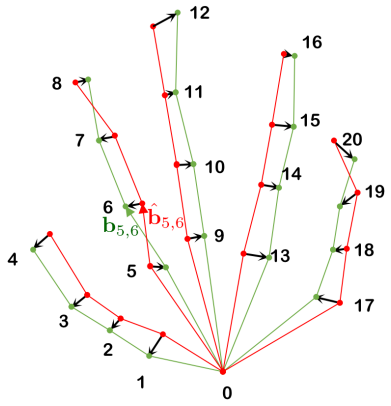

Figure 3: **Illustration of the residual between the ground truth hand pose (marked in green) and the predicted one (marked in red).** Each hand pose has 21 key joints. We denote a bone vector connecting two key joints i and j by $\mathbf{b}_{i,j}$, such as $\mathbf{b}_{5,6}$ in the figure.

The Hand Model Module Given an input monocular image, this module aims to generate an initial estimation of 3D hand pose $\tilde{\mathbf{P}}$ as a prior. A hand model is able to represent both hand shape and pose with a few parameters, which is thus a suitable prior for hand pose estimation.

We first predict parameters of the hand model. Specifically, we crop and resize the input image to a salient region of the hand, which is fed into the ResNet-50 network (He et al. 2016) to extract features for the construction of the latent code $\mathbf{z}$, *i.e.*, parameters of the hand model. Then, we employ a modified MANO hand model (Romero, Tzionas, and Black 2017), which is based on the SMPL model (Loper et al. 2015) for human bodies. The MANO hand model is a deformable hand mesh model with two vectors θ and β contained in the latent code $\mathbf{z}$ as the input, which control the pose and shape of the generated hand respectively. We modify the default setting of $\{\theta, \beta\}$ from $\{10, 45\}$ to $\{10, 8\}$ for reduced computation complexity. Also, note that, while Boukhayma *et al.* (Boukhayma, Bem, and Torr 2019) create a synthetic dataset to pre-train the ResNet-50 so as to estimate parameters of MANO, we do not resort to any extra dataset. The output of the MANO hand model includes a hand pose $\mathcal{P}(\theta, \beta)$.

Additionally, we need to position the pose $\mathcal{P}(\theta, \beta)$ in a camera coordinate system so as to acquire the 3D coordinates of each point in the hand pose. We project $\mathcal{P}(\theta, \beta)$ to the 3D space via three parameters that model the camera coordinate system: 1) a 3D rotation parameter $\mathbf{c_r} \in \mathbb{R}^3$; 2) a 3D translation parameter $\mathbf{c_t} \in \mathbb{R}^3$; and 3) a scale parameter $\mathbf{c_s} \in \mathbb{R}$. The camera parameters are estimated by the aforementioned ResNet-50 network.

We formulate the complete hand model as:

$$\tilde{\mathbf{P}}(\theta, \beta, \mathbf{c_r}, \mathbf{c_t}, \mathbf{c_s}) = \mathbf{c_s} * R(\mathcal{P}(\theta, \beta), \mathbf{c_r}) + \mathbf{c_t}, \quad (4)$$

where R is a rotation function. The acquired initial estimation $\tilde{\mathbf{P}}$ serves as a prior pose for the subsequent deformation learning model.

The Deformation Learning Module This module aims at accurate graph representation learning for hand pose estimation, which is conditional on the prior and under the supervision of the input image and ground truth pose as in (1). In particular, conditioned on the prior $\tilde{\mathbf{P}}$, we learn the structural *deformation* in $\tilde{\mathbf{P}}$ instead of the holistic hand pose.

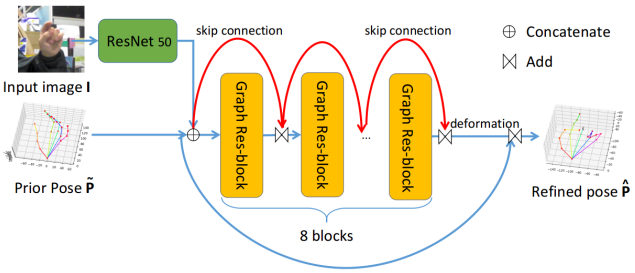

Figure 4: Architecture of the deformation learning module in our generator.

We first construct an unweighted graph over $\tilde{\mathbf{P}}$, where the irregularly sampled key points (*i.e.*, joints) on the hand are projected onto nodes. The graph signal on each node is the concatenation of the global feature vector of the input image and the 3-dimensional coordinate vector of each joint in the input prior pose. Nodes are connected if they represent adjacent key points of the hand, where the adjacency relations follow the human hand structure as presented in Fig. 3, leading to an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$.

Based on the graph representation $\mathbf{A}$, the finally refined pose is

$$\hat{\mathbf{P}} = \tilde{\mathbf{P}} + \text{GCN}(\tilde{\mathbf{P}} \oplus \mathbf{F}, \mathbf{A}), \quad (5)$$

where $\mathbf{F} \in \mathbb{R}^{N \times F}$ denotes the F-dimensional global feature vector of the image repeated N times, and $\oplus$ denotes the feature-wise concatenation operation. $\text{GCN}(\tilde{\mathbf{P}} \oplus \mathbf{F}, \mathbf{A})$ represents the learned deformation between the prior $\tilde{\mathbf{P}}$ and the ground truth. The sum of the prior pose $\tilde{\mathbf{P}}$ and its deformation thus leads to the refined hand pose.

Specifically, we first extract features from the RGB image $\mathbf{I}$ to facilitate the pose refinement. We follow the ResNet-50 architecture (He et al. 2016) to extract 2D features $\mathbf{F}$ from $\mathbf{I}$. Next, we learn the structural deformation in the hand pose via residual graph convolution. Leveraging on the idea of the recent ResGCN (Li et al. 2019) that shows graph residual learning enables deeper graph convolution networks and better feature learning, we design a *Graph Res-block* to learn the deformation of the prior pose. Specifically, we employ the efficient GCN (Kipf and Welling 2017) as the basic unit of the Graph Res-block, which essentially propagates information across adjacent nodes to learn higher-level features. Each Graph Res-block consists of two GCN layers as well as two normalization layers that enable higher learning rate without vanishing or exploding gradients. Furthermore, we introduce residual skip connections for all the Graph Res-blocks in order to accelerate the speed of convergence and avoid the gradient vanishment.

Let $\mathbf{X}^l$ denote the input of the l-th Graph Res-block, then the output of the l-th Graph Res-block takes the form

$$\mathbf{X}^{l+1} = N\left(g(N(g(\mathbf{X}^l, \mathbf{A})), \mathbf{A})\right) + \text{skip}(\mathbf{X}^l), \quad (6)$$

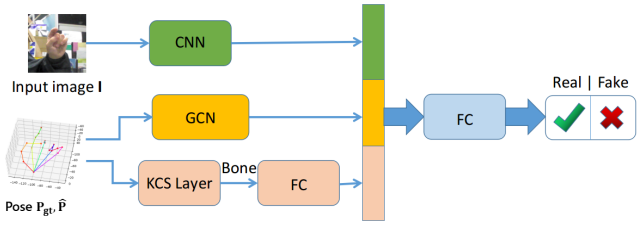

Figure 5: Architecture of the conditional discriminator.

where $g(\cdot)$ represents a single GCN layer as in (Kipf and Welling 2017), $N(\cdot)$ represents a single normalization layer, and skip$(\cdot)$ denotes the skip connection which is a GCN layer to match the dimension of the two terms in (6). We then stack several layers of Graph Res-blocks to learn the deformation of the prior pose, as demonstrated in Fig. 4.

3.3 The Proposed Conditional Discriminator $\mathbb{D}$

In the adversarial training stage, while we learn the generator to predict hand poses which are indistinguishable to the discriminator, the discriminator attempts to distinguish real samples from fake ones, *i.e.*, the predicted hand poses. In particular, given the input image $\mathbf{I}$, we designate a *conditional* discriminator conditioned on $\mathbf{I}$.

A simple architecture of a discriminator is a fully-connected (FC) network with the hand pose as input, which however has two shortcomings: 1) the relation between the RGB image and inferred hand pose is neglected; 2) structural properties of the hand pose are not taken into account explicitly. Instead, inspired by the multi-source architecture in (Yang et al. 2018), we design a conditional multi-source discriminator with three inputs to address the aforementioned issues. As illustrated in Fig. 5, the inputs include: 1) features of the input monocular image; 2) features of the refined hand pose $\hat{\mathbf{P}}$ or the ground truth pose $\mathbf{P}_{\text{gt}}$; 3) features of bones via the KCS layer as in (Wandt, Ackermann, and Rosenhahn 2017), which computes the bone matrix from $\hat{\mathbf{P}}$ or $\mathbf{P}_{\text{gt}}$ via a simple matrix multiplication. The bone features contain prominent structural information such as the length and direction of bones, thus characterizing the hand structure accurately.

The loss function of the conditional discriminator follows the definition of the Wasserstein loss (Arjovsky, Chintala, and Bottou 2017) conditioned on the input image $\mathbf{I}$:

$$\mathcal{L}_{\text{Wass}} = -\mathbb{E}_{\mathbf{P}_{\text{gt}}\sim p_{data}(\mathbf{P}_{\text{gt}})}\mathbb{D}(\mathbf{P}_{\text{gt}}|\mathbf{I}) + \mathbb{E}_{\hat{\mathbf{P}}\sim p(\hat{\mathbf{P}})}\mathbb{D}(\hat{\mathbf{P}}|\mathbf{I}), \quad (7)$$

where $\mathbb{D}$ takes the generated (fake) pose $\hat{\mathbf{P}}$ and ground-truth pose $\mathbf{P}_{\text{gt}}$ as input, $\mathbf{P}_{\text{gt}}$ is a sample following the ground-truth pose distribution $p_{data}(\mathbf{P}_{\text{gt}})$ and $\hat{\mathbf{P}}$ is a sample from the refined pose distribution $p(\hat{\mathbf{P}})$.

Specifically, we employ a CNN to extract features of the input monocular image, a GCN to learn the representation of the refined pose or the ground truth pose, and one FC layer to capture the features of bone structures computed from the hand pose. Besides, the architecture of our multi-source discriminator is based on SNGAN (Miyato et al. 2018) with spectral normalization layers.

3.4 The Proposed Bone-Constrained Loss Functions

As presented in (3), we have two types of loss functions for the MAP estimation of hand pose. We employ the commonly adopted average Euclidean distance in the coordinates of joints of 3D hand pose $\mathcal{L}_{\text{pose}}$ (Ge et al. 2019) as well as two proposed bone-constrained metrics as $d_1(\cdot)$ to measure the distortion of the estimated 3D hand pose compared to the ground truth, and apply the commonly used average Euclidean distance in the coordinates of joints of projected 2D hand pose $\mathcal{L}_{\text{proj}}$ (Ge et al. 2019) as $d_2(\cdot)$ to measure the distance between the estimation and the 2D image.

Since $\mathcal{L}_{\text{pose}}$ and $\mathcal{L}_{\text{proj}}$ cannot capture the structural properties of hand pose explicitly, we propose two novel bone-constrained loss functions to characterize the length and direction of each bone.

As illustrated in Fig. 3, we first define a bone vector $\mathbf{b}_{i,j} \in \mathbb{R}^{3\times 1}$ between hand joint i and j as

$$\mathbf{b}_{i,j} = \mathbf{j}_i - \mathbf{j}_j, \quad (8)$$

where $\mathbf{j}_i, \mathbf{j}_j \in \mathbb{R}^{3\times 1}$ are the coordinates of joint i and j respectively.

The first bone-constrained loss $\mathcal{L}_{\text{len}}$ quantifies the distance in *bone length* between the ground truth hand and its estimate, which we define as

$$\mathcal{L}_{\text{len}} = \sum_{i,j} \left| ||\mathbf{b}_{i,j}||_2 - ||\hat{\mathbf{b}}_{i,j}||_2 \right|, \quad (9)$$

where $\mathbf{b}_{i,j}$ and $\hat{\mathbf{b}}_{i,j}$ are the bone vectors of the ground truth and the predicted bone respectively.

The second bone-constrained loss $\mathcal{L}_{\text{dir}}$ measures the deviation in the *direction of bones*:

$$\mathcal{L}_{\text{dir}} = \sum_{i,j} \left| \left| \mathbf{b}_{i,j}/||\mathbf{b}_{i,j}||_2 - \hat{\mathbf{b}}_{i,j}/||\hat{\mathbf{b}}_{i,j}||_2 \right| \right|_2. \quad (10)$$

This is motivated by the fact that small loss in joints sometimes may not reflect large distortion in hand pose. Taking joints $\mathbf{j}_5$ and $\mathbf{j}_6$ in Fig. 3 as an example, the distance between the ground truth joints and predicted ones is trivial. However, it is obvious that the orientation of the predicted bone $\hat{\mathbf{b}}_{5,6}$ significantly deviates from the ground truth $\mathbf{b}_{5,6}$. This distortion in hand structure is well captured by our proposed loss in the bone direction $\mathcal{L}_{\text{dir}}$.

Besides, as we adopt the framework of adversarial learning, we also introduce the Wasserstein loss $\mathcal{L}_{\text{Wass}}$ in (7) into the loss function for adversarial training. Hence, the overall loss function $\mathcal{L}$ is

$$\mathcal{L} = \mathcal{L}_{\text{pose}} + \lambda_{\text{proj}}\mathcal{L}_{\text{proj}} + \lambda_{\text{len}}\mathcal{L}_{\text{len}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{Wass}}\mathcal{L}_{\text{Wass}}, \quad (11)$$

where λ_{proj}, λ_{len}, λ_{dir} and λ_{Wass} are hyperparameters for the trade-off among these losses. In accordance with (3), $d_1 = \mathcal{L}_{\text{pose}} + \lambda_{\text{len}}\mathcal{L}_{\text{len}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}}$, and $d_2 = \lambda_{\text{proj}}\mathcal{L}_{\text{proj}}$.

Stage	hand model	deformation	discriminator	STB	RHD	EGODEXTER
I	✓			24.15	83.37	52.32
II	✓	✓		5.12	15.84	43.26
III	✓	✓	✓	**3.97**	**12.40**	**34.98**

Table 1: The performance of different stages in our model on three datasets (measured in 3D Euclidean distance (mm)).

	GCN Deformation	FC Deformation	Discriminator	STB	RHD	EGODEXTER
1		✓		15.11	37.59	52.34
2	✓			5.12	15.84	40.12
3		✓	✓	10.23	25.15	44.23
4	✓		✓	**3.97**	**12.40**	**34.98**

Table 2: Ablation studies on the Deformation Learning Module, with comparison between the Deformation Learning Module and the simple FC Refinement Module in 3D Euclidean distance (mm).

4 Experimental Results

4.1 Datasets and Metrics

Datasets We evaluate our approach on five public datasets: Stereo Hand Pose Tracking Benchmark (STB) (Zhang et al. 2016b), the Rendered Hand Pose Dataset (RHD) (Zimmermann and Brox 2017b), EGODEXTER (Mueller et al. 2017), MPII+NZSL (Simon et al. 2017) and DEXTER+OBJECT (Mueller et al. 2017). We use STB, RHD and EGODEXTER for ablation studies and compare with state-of-the-art methods on all the five datasets.

STB is a real-world dataset with image resolution of 640×320. Following (Zimmermann and Brox 2017b), we split the 18,000 images into 15,000 training samples and 3,000 test samples. Besides, to make the definition of joints consistent, we move the location of the root joint from the palm center to the wrist following (Ge et al. 2019). RHD is a more challenging synthetic dataset with image resolution of 320×320, which is built upon 20 different characters performing 39 actions. MPII+NZSL is a real-world dataset containing images from YouTube videos. This dataset only provides 2D annotations. DEXTER+OBJECT dataset shows interactions of an actor's hand with a cuboid object from a third person view. EGODEXTER dataset displays a hand from an egocentric view interacting with various objects.

Metrics We evaluate the performance of 3D hand pose estimation with two metrics: (i) pose error, which takes the average location error in Euclidean distance between the estimated 3D joints and the ground truth; (ii) percentage of correct key points (PCK), which is the percentage of correct key points whose error in Euclidean distance is below a threshold.

4.2 Implementation Details

In our experiments, we first pretrain the hand model module and then train the entire network end-to-end. In particular, the training process can be divided into three stages.

Stage I. We pretrain the hand model module, which is randomly initialized and trained for 100 epochs using the Adam optimizer with learning rate 0.001. Then, we freeze the parameters of this stage to evaluate the effectiveness of the deformation learning module.

Stage II. We train the generator $\mathbb{G}$ end-to-end without the discriminator $\mathbb{D}$. In $\mathbb{G}$, the hand model module is initialized with the trained model in the first stage and the deformation learning module is randomly initialized. $\mathbb{G}$ is then trained with 100 epochs using the Adam optimizer with learning rate 0.0001.

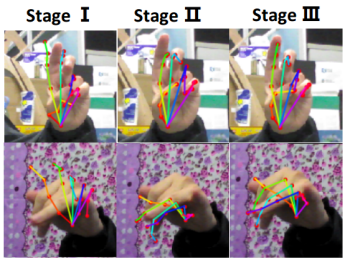

Figure 6: Qualitative results of different stages in our model.

Stage III. We adopt the framework of SNGAN (Miyato et al. 2018) for the conditional adversarial training, and train our model end-to-end. $\mathbb{G}$ and $\mathbb{D}$ are trained with 100 epochs using the Adam optimizer with learning rate 0.0001.

Regarding the hyper-parameters in (11), we set $\lambda_{\text{len}} = 0.01, \lambda_{\text{dir}} = 0.1, \lambda_{\text{proj}} = 0.1, \lambda_{\text{Wass}} = 0.01$.

4.3 Ablation Studies

We perform ablation studies on the performance of different stages, the deformation learning module, the discriminator and loss functions. Due to the page limit, we present all the results in 3D Euclidean distance (mm). Please refer to the supplementary material for the results measured in 3D PCK.

On different stages. We present the results of three training stages in average 3D Euclidean distance, as listed in Tab. 1. The performance of **Stage II** significantly outperforms **Stage I**, which demonstrates that the proposed deformation learning module plays the most critical role in our model. The adversarial training scheme (**Stage III**) further improves the result, by learning a real distribution of the 3D hand pose. We also show visual results of our method at different stages in Fig. 6. We see that Stage I estimates a coarse hand pose from the MANO hand model as a prior pose, while Stage II refines the structure of the prior pose significantly. Finally, Stage III generates more realistic hand poses via conditional adversarial learning.

On the deformation learning module. We compare the deformation learning module with a simple fully-connected deformation learning module (FC Deformation Module) to refine the prior pose. We train the deformation learning modules in different experimental settings: 1) without our discriminator, *i.e.*, without adversarial learning; and 2) with our discriminator. As presented in Tab. 2, the GCN deformation learning module leads to significant gain over the simple FC deformation module on both datasets in different settings, thus validating the superiority of the proposed deformation learning module.

On the conditional discriminator. We compare with a single-source discriminator which only takes the 3D hand pose as the input. As presented in Tab. 3, the multi-source discriminator outperforms the single-source one on both datasets, which gives credits to exploring the structure of hand bones and the relation between the image and pose.

	Deformation Learning	Multi-source	Single-source	STB	RHD	EGODEXTER
1	✓	✓		**3.97**	**12.40**	**34.98**
2	✓		✓	4.54	15.10	37.46

Table 3: Ablation studies on the discriminator (3D Euclidean distance (mm)).

	$\mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{proj}}$	$\mathcal{L}_{\text{len}}$	$\mathcal{L}_{\text{dir}}$	STB			RHD		
				Stage I	Stage II	Stage III	Stage I	Stage II	Stage III
1	✓			32.75	9.11	5.35	99.24	25.96	15.07
2	✓	✓		30.32	8.00	5.02	95.19	22.96	14.76
3	✓		✓	27.65	6.91	5.00	89.76	21.63	14.01
4	✓	✓	✓	24.15	5.12	**3.97**	83.37	15.84	**12.40**

Table 4: Ablation studies on the proposed bone-constrained loss functions at three stages.

On loss functions. We also evaluate the proposed bone-constrained loss functions $\mathcal{L}_{\text{len}}$ and $\mathcal{L}_{\text{dir}}$ separately. We train the network with different combinations of loss functions on the STB and RHD datasets in three stages respectively. As reported in Tab. 4, the network trained with our proposed bone-constrained loss functions performs better in all the three stages on both datasets. We also notice that $\mathcal{L}_{dir}$ plays a more significant role compared to $\mathcal{L}_{\text{len}}$. This gives credits to the constraint on the orientation of bones that explicitly takes structural properties of hands into consideration. Further, we demonstrate the visual comparison of estimated poses with and without bone-constrained losses in Fig. 7. The estimated pose may have unnatural distortion in the direction of bones in the absence of the bone-constrained loss functions, *e.g.*, the little finger in the first row and the thumb in the second row. In contrast, our results exhibit natural and accurate structure in the orientation of bones with the proposed bone constraints enforced.

4.4 Experimental Results

	STB	RHD	MPII+ZNSL(px)	Dexter+Object	EgoDexter
(Ge et al. 2019)	6.37	15.33	-	-	-
(Boukhayma, Bem, and Torr 2019)	9.76	-	18.95	25.53	45.33
(Spurr et al. 2018)	8.56	19.73	-	40.20	56.92
(Zimmermann and Brox 2017b)	-	-	59.40	34.75	52.77
Ours	**3.97**	**12.40**	**9.87**	**16.12**	**34.98**

Table 5: Comparison with state-of-the-art methods on the five datasets. Note that MPII+ZNSL only provides 2D annotation, thus we employ the 2D distance (px) metric on this dataset.

We compare our method with competitive 3D hand pose estimation approaches on the five datasets. On the relatively simple STB dataset, as shown in Fig. 8, we compare with latest methods (Cai et al. 2018; Iqbal et al. 2018; Boukhayma, Bem, and Torr 2019; Ge et al. 2019; Mueller et al. 2018a; Zimmermann and Brox 2017a; Mueller et al. 2018b). Our paradigm outperforms the state-of-the-art (Ge et al. 2019), which closely reaches the upper bound 1.0 of 3D PCK at all the error thresholds. Also, we list the results in 3D Euclidean distance for comparison with the state-of-the-arts in Tab. 5. Compared to these works which directly reconstruct the 3D hand pose (Ge et al. 2019; Boukhayma, Bem, and Torr 2019; Cai et al. 2018), our method performs much better mainly due to the proposed regularized graph representation learning and conditional adversarial learning.

Moreover, we demonstrate some qualitative results of our

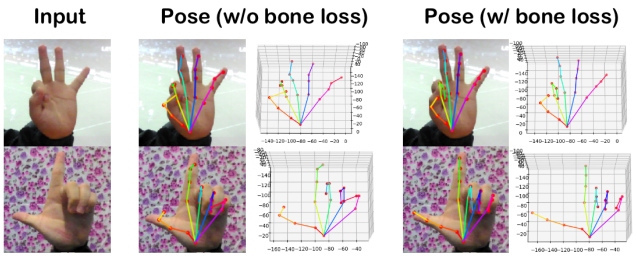

Figure 7: Qualitative results for the evaluation of the proposed bone-constrained loss functions.

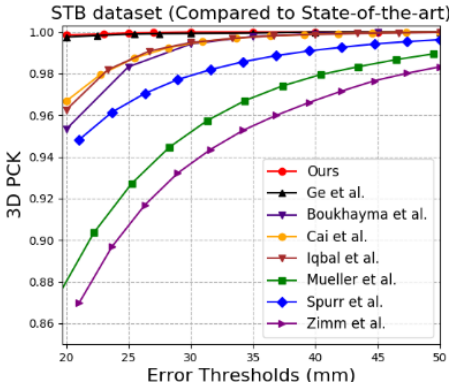

Figure 8: Comparison with state-of-the-art methods on the STB dataset.

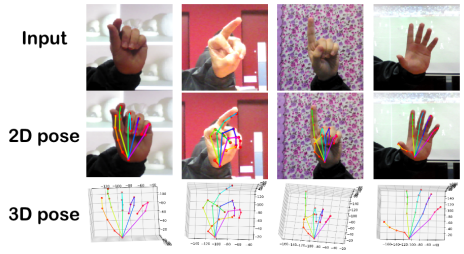

Figure 9: Qualitative results of our proposed network on the STB dataset. The 2D pose is projected from the 3D pose.

3D hand pose estimation in Fig. 9. The generated poses are accurate and natural even in case of severe self-occlusions, as shown in the first three columns of Fig. 9. This validates the effectiveness of the proposed paradigm. We show the qualitative results and PCK results on the other four datasets in the supplementary material.

5 Conclusion

In this paper, we propose regularized graph representation learning under a conditional adversarial learning framework for 3D hand pose estimation from a monocular image. Based on the MAP estimation formulation, we take an initial estimation of hand pose as prior pose, and further learn the structural deformation in the prior pose via residual graph convolution. Also, we propose two bone-constrained loss functions to enforce constraints on the bone structures explicitly. Extensive experiments demonstrate the superiority of the proposed method.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Generative Adversarial Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 214–223. International Convention Centre, Sydney, Australia: PMLR.

Athitsos, V.; and Sclaroff, S. 2003. Estimating 3D Hand Pose From a Cluttered Image. In *IEEE Computer Society Conference on Computer Vision & Pattern Recognition*.

Baek, S.; Kim, K. I.; and Kim, T.-K. 2019. Pushing the Envelope for RGB-Based Dense 3D Hand Pose Estimation via Neural Rendering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Boukhayma, A.; Bem, R. d.; and Torr, P. H. 2019. 3D Hand Shape and Pose From Images in the Wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cai, Y.; Ge, L.; Cai, J.; and Yuan, J. 2018. Weakly-supervised 3D Hand Pose Estimation from Monocular RGB Images. In *The European Conference on Computer Vision (ECCV)*.

Campos, T. E. D.; and Murray, D. W. 2006. Regression-based Hand Pose Estimation from Multiple Cameras. In *IEEE Computer Society Conference on Computer Vision & Pattern Recognition*.

Choi, C. 2016. DeepHand: Robust Hand Pose Estimation by Completing a Matrix Imputed with Deep Features. In *Computer Vision & Pattern Recognition*.

De, L. G. M.; Fleet, D. J.; and Paragios, N. 2011. Model-Based 3D Hand Pose Estimation from Monocular Video. *IEEE Trans Pattern Anal Mach Intell* 33(9): 1793–1805.

Doosti, B.; Naha, S.; Mirbagheri, M.; and Crandall, D. J. 2020. HOPE-Net: A Graph-Based Model for Hand-Object Pose Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fitzgibbon, A. 2015. Accurate, Robust, and Flexible Real-time Hand Tracking. *Inproceedings* 3633–3642.

Ge, L.; Cai, Y.; Weng, J.; and Yuan, J. 2018. Hand PointNet: 3D Hand Pose Estimation Using Point Sets. 8417–8426. doi:10.1109/CVPR.2018.00878.

Ge, L.; Liang, H.; Yuan, J.; and Thalmann, D. 2016. Robust 3D Hand Pose Estimation in Single Depth Images: From Single-View CNN to Multi-View CNNs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ge, L.; Ren, Z.; Li, Y.; Xue, Z.; Wang, Y.; Cai, J.; and Yuan, J. 2019. 3D Hand Shape and Pose Estimation From a Single RGB Image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hui, L.; Yuan, J.; Lee, J.; Ge, L.; and Thalmann, D. 2017. Hough Forest With Optimized Leaves for Global Hand Pose Estimation With Arbitrary Postures. *IEEE Transactions on Cybernetics* PP(99): 1–15.

Hürst, W.; and van Wezel, C. 2011. Gesture-based interaction via finger tracking for mobile augmented reality. *Multimedia Tools and Applications* 62: 233–258.

Iqbal, U.; Molchanov, P.; Breuel Juergen Gall, T.; and Kautz, J. 2018. Hand Pose Estimation via Latent 2.5D Heatmap Regression. In *The European Conference on Computer Vision (ECCV)*.

Khamis, S.; Taylor, J.; Shotton, J.; Keskin, C.; Izadi, S.; and Fitzgibbon, A. 2015. Learning an efficient model of hand shape variation from depth images. In *IEEE Conference on Computer Vision & Pattern Recognition*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

Kulon, D.; Wang, H.; Güler, R. A.; Bronstein, M. M.; and Zafeiriou, S. 2019. Single Image 3D Hand Reconstruction with Mesh Convolutions. In *BMVC*.

Li, G.; Muller, M.; Thabet, A.; and Ghanem, B. 2019. DeepGCNs: Can GCNs Go As Deep As CNNs? In *The IEEE International Conference on Computer Vision (ICCV)*.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-person Linear Model. *ACM Trans. Graph.* 34(6): 248:1–248:16. ISSN 0730-0301. doi:10.1145/2816795.2818013. URL http://doi.acm.org/10.1145/2816795.2818013.

Malik, J.; Elhayek, A.; Nunnari, F.; Varanasi, K.; and Stricker, D. 2018. DeepHPS: End-to-end Estimation of 3D Hand Pose and Shape by Learning from Synthetic Depth. In *2018 International Conference on 3D Vision (3DV)*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=B1QRgziT-.

Mueller, F.; Bernard, F.; Sotnychenko, O.; Mehta, D.; Sridhar, S.; Casas, D.; and Theobalt, C. 2018a. GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. URL https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/.

Mueller, F.; Bernard, F.; Sotnychenko, O.; Mehta, D.; Sridhar, S.; Casas, D.; and Theobalt, C. 2018b. GANerated Hands for Real-Time 3D Hand Tracking From Monocular RGB. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mueller, F.; Mehta, D.; Sotnychenko, O.; Sridhar, S.; Casas, D.; and Theobalt, C. 2017. Real-time Hand Tracking under Occlusion from an Egocentric RGB-D Sensor. In *Proceedings of International Conference on Computer Vision (ICCV)*. URL https://handtracker.mpi-inf.mpg.de/projects/OccludedHands/.

Oberweger, M.; and Lepetit, V. 2018. DeepPrior++: Improving Fast and Accurate 3D Hand Pose Estimation. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*.

Oikonomidis, I.; Kyriazis, N.; and Argyros, A. 2011. Efficient model-based 3D tracking of hand articulations using Kinect. volume 1. doi:10.5244/C.25.101.

Oikonomidis, I.; Kyriazis, N.; and Argyros, A. A. 2010. Markerless and Efficient 26-DOF Hand Pose Recovery. In *Computer Vision-accv -asian Conference on Computer Vision*.

Panteleris, P.; and Argyros, A. A. 2017. Back to RGB: 3D tracking of hands and hand-object interactions based on short-baseline stereo. *CoRR* abs/1705.05301. URL http://arxiv.org/abs/1705.05301.

Piumsomboon, T.; Clark, A.; Billinghurst, M.; and Cockburn, A. 2013. User-Defined Gestures for Augmented Reality. In Kotzé, P.; Marsden, G.; Lindgaard, G.; Wesson, J.; and Winckler, M., eds., *Human-Computer Interaction – INTERACT 2013*, 282–299. Berlin, Heidelberg: Springer Berlin Heidelberg.

Rehg, J. M.; and Kanade, T. 1994. Visual tracking of high DOF articulated structures: An application to human hand tracking. In *Proc of Third European Conference on Computer Vision*.

Rehg, J. M.; and Kanade, T. 2002. DigitEyes: vision-based hand tracking for human-computer interaction. In *IEEE Workshop on Motion of Non-rigid & Articulated Objects*.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Trans. Graph.* 36(6): 245:1–245:17. ISSN 0730-0301. doi:10.1145/3130800.3130883. URL http://doi.acm.org/10.1145/3130800.3130883.

Simon, T.; Joo, H.; Matthews, I.; and Sheikh, Y. 2017. Hand Keypoint Detection in Single Images Using Multiview Bootstrapping. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Spurr, A.; Song, J.; Park, S.; and Hilliges, O. 2018. Cross-modal Deep Variational Hand Pose Estimation. *CoRR* abs/1803.11404. URL http://arxiv.org/abs/1803.11404.

Sridhar, S.; Rhodin, H.; Seidel, H. P.; Oulasvirta, A.; and Theobalt, C. 2014. Real-time hand tracking using a sum of anisotropic gaussians model. In *International Conference on 3D Vision*.

Stenger, B.; Thayananthan, A.; Torr, P. H. S.; and Cipolla, R. 2006. Model-based hand tracking using a hierarchical Bayesian filter. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(9): 1372–1384. doi:10.1109/TPAMI.2006.189.

Tang, D.; Yu, T. H.; and Kim, T. K. 2013. Real-Time Articulated Hand Pose Estimation Using Semi-supervised Transductive Regression Forests. In *IEEE International Conference on Computer Vision*.

Tkach, A.; Pauly, M.; and Tagliasacchi, A. 2016. Sphere-meshes for Real-time Hand Modeling and Tracking. *ACM Trans. Graph.* 35(6): 222:1–222:11. ISSN 0730-0301. doi:10.1145/2980179.2980226. URL http://doi.acm.org/10.1145/2980179.2980226.

Wandt, B.; Ackermann, H.; and Rosenhahn, B. 2017. A Kinematic Chain Space for Monocular Motion Capture .

Wu, X.; Finnegan, D.; O'Neill, E.; and Yang, Y.-L. 2018. HandMap: Robust Hand Pose Estimation via Intermediate Dense Guidance Map Supervision. In *The European Conference on Computer Vision (ECCV)*.

Xiao, S.; Wei, Y.; Shuang, L.; Tang, X.; and Jian, S. 2015. Cascaded hand pose regression. In *IEEE Conference on Computer Vision & Pattern Recognition*.

Yang, L.; and Yao, A. 2019. Disentangling Latent Hands for Image Synthesis and Pose Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, W.; Ouyang, W.; Wang, X.; Ren, J.; Li, H.; and Wang, X. 2018. 3D Human Pose Estimation in the Wild by Adversarial Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ying, W.; and Huang, T. S. 2002. Hand modeling, analysis and recognition. *IEEE Signal Processing Magazine* 18(3): 51–60.

Ying, W.; John, L.; and Huang, T. S. 2005. Analyzing and capturing articulated hand motion in image sequences. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 27(12): 1910–1922.

Yuan, S.; Garcia-Hernando, G.; Stenger, B.; Moon, G.; Chang, J.; Lee, K.; Molchanov, P.; Kautz, J.; Honari, S.; Ge, L.; Yuan, J.; Chen, X.; Wang, G.; Yang, F.; Akiyama, K.; Wu, Y.; Wan, Q.; Madadi, M.; Escalera, S.; and Kim, T.-K. 2018. Depth-Based 3D Hand Pose Estimation: From Current Achievements to Future Goals. 2636–2645. doi:10.1109/CVPR.2018.00279.

Zhang, J.; Jiao, J.; Chen, M.; Qu, L.; Xu, X.; and Yang, Q. 2016a. 3D Hand Pose Tracking and Estimation Using Stereo Matching .

Zhang, J.; Jiao, J.; Chen, M.; Qu, L.; Xu, X.; and Yang, Q. 2016b. 3D Hand Pose Tracking and Estimation Using Stereo Matching. *ArXiv* abs/1610.07214.

Zhou, Y.; Lu, J.; Du, K.; Lin, X.; Sun, Y.; and Ma, X. 2018. HBE: Hand Branch Ensemble Network for Real-time 3D Hand Pose Estimation. In *The European Conference on Computer Vision (ECCV)*.

Zimmermann, C.; and Brox, T. 2017a. Learning to Estimate 3D Hand Pose from Single RGB Images. 4913–4921. doi:10.1109/ICCV.2017.525.

Zimmermann, C.; and Brox, T. 2017b. Learning to Estimate 3D Hand Pose From Single RGB Images. In *The IEEE International Conference on Computer Vision (ICCV)*.