Aligning Latent Spaces for 3D Hand Pose Estimation

Linlin Yang*[1], Shile Li*[2], Dongheui Lee[2,3], Angela Yao[4]
*Equal contribution
[1]University of Bonn, Germany [2] Technical University of Munich, Germany
[3]German Aerospace Center, Germany [4]National University of Singapore, Singapore

Abstract

Hand pose estimation from monocular RGB inputs is a highly challenging task. Many previous works for monocular settings only used RGB information for training despite the availability of corresponding data in other modalities such as depth maps. In this work, we propose to learn a joint latent representation that leverages other modalities as weak labels to improve RGB-based hand pose estimation. By design, our architecture is highly flexible in embedding various diverse modalities such as heat maps, depth maps and point clouds. In particular, we find that encoding and decoding the point cloud of the hand surface can improve the quality of the joint latent representations. Experiments show that with the aid of other modalities during training, our proposed method boosts the accuracy of RGB-based hand pose estimation systems and significantly outperforms state-of-the-art on two public benchmarks.

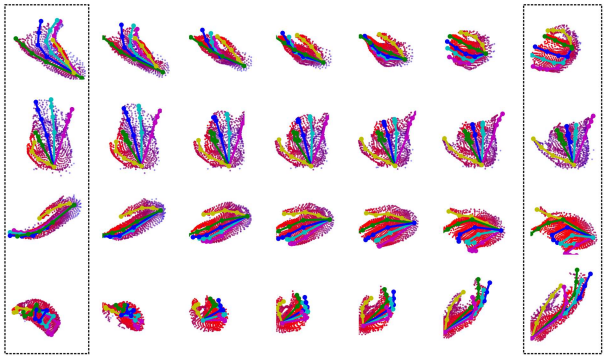

Figure 1: Latent space interpolation. The far left and far right columns (dashed boxes) are generated poses and point clouds from monocular RGB images sampled from the training data. Other columns are generated from linear interpolations on the latent space. The smoothness and consistency imply that different cross-modal latent spaces can be embedded and aligned into one shared latent space.

1. Introduction

Hand pose estimation plays an important role in areas such as human activity analysis, human computer interaction, and robotics. Depth-based 3D hand pose estimation methods are now highly accurate [25, 10, 28] largely due to advancements from deep learning. Despite commodity depth sensors being more commonplace, high-quality depth maps can still only be captured indoors, thereby limiting the environments in which depth-based methods can be deployed. Furthermore, simple RGB cameras, as well as existing RGB footage are still far more ubiquitous than depth cameras and depth data. As such, there is still a need for accurate RGB-based 3D hand pose estimation methods, especially from monocular viewpoints.

To tackle the ambiguities associated with monocular RGB inputs, previous works have relied on large amounts of training data [31, 12]. Gains from purely increasing dataset size tend to saturate, because it is very difficult to obtain accurate ground truth labels, *i.e.* 3D hand poses. Annotating 3D hand joint positions accurately is a difficult task and there is often little consensus between human annotators [20]. While several methods have been developed to generate RGB images [12], there still exists a large domain gap between synthesized and real-world data, limiting the utility of synthetic data.

Even though accurate ground truth for RGB data is hard to collect, there exists plenty of unlabelled RGB-D hand data which can be leveraged together with labelled depth maps. Cai *et al.* [2] first proposed the use of labelled depth maps as regularizers to boost RGB-based methods. Yang *et al.* [27] introduced a disentangled representation so that viewpoint can be used as a weak label. Inspired by these works, we aim to leverage multiple modalities as weak labels for enhancing RGB-based hand pose estimation.

In this paper, we consider different modalities of hand data (*e.g.* RGB images, depth maps, point clouds, 3D poses, heat maps and segmentation masks) and formulate RGB-based hand pose estimation as a cross-modal inference problem. In particular, we propose the use of a multi-modal variational autoencoder (VAE). VAEs are an attrac-

tive class of deep generative models which can be learned on large-scale, high-dimensional datasets. They have been shown to capture highly complex relationships across multiple modalities [21, 24, 26] and have also been applied to RGB-based pose estimation in the past [19, 27]. However, both [19] and [27] learn a single shared latent space and as a result must compromise on pose reconstruction accuracy.

In this work, we propose to align latent space from individual modalities. More specifically, we derive different objectives for three diverse modalities, namely 3D poses, point clouds, and heat maps, and show two different ways to aligning their associated hand latent spaces. While such a solution may appear less elegant than learning one shared latent space directly, it is has several practical advantages. First and foremost, it is much faster to converge and results in a well-structured latent space; in comparison, the multimodal shared latent space of [19] tends to fluctuate as one draws data from the multiple modalities. Additionally, the learning scheme through alignment offers more flexibility in working with non-corresponding data and also weak supervision. The resulting latent representation allows for estimating highly accurate hand poses and synthesizing realistic-looking point clouds of the hand surface, all from monocular RGB images (See Fig. 1).

The main contributions of this paper are as follows:

- We formulate RGB-based hand pose estimation as a multi-modal learning, cross-modal inference problem and propose three strategies for learning from different hand inputs of various modalities.
- We explore non-conventional inputs such as point clouds and heat maps for learning the latent hand space and show how they can be leveraged for improving the accuracy of an RGB-based hand pose estimation system. A side product of our framework is that we can synthesize realistic-looking point clouds of the hand from RGB images.
- By evaluating on two publicly available benchmarks, we show that our proposed framework makes full use of auxiliary modalities during training and boosts the accuracy of RGB pose estimates. Our estimated poses surpass state-of-the-art methods on monocular RGB-based hand pose estimation, including a whopping 19% improvement on the challenging RHD dataset [31]

2. Related Works

One way to categorize hand pose estimation approaches is according to either generative or discriminative methods. Generative methods employ a hand model and use optimization to fit the hand model to the observations [17, 14, 22]. They usually require a good initialization; otherwise they are susceptible to getting stuck in local minima. Discriminative methods learn a direct mapping from visual observations to hand poses [23, 27, 10, 13, 31, 2]. Thanks to large-scale annotated datasets [31, 29, 23], deep learning-based discriminative methods have shown very strong performance in the hand pose estimation task.

In particular, works using depth or 3D data as input are the most accurate. Oberweger *et al.* [13] use 2D CNNs to regress the hand pose from depth images, using a bottleneck layer to regularize the pose prediction to a certain prior distribution. Moon *et al.* [11] use 3D voxels as input and regress the hand pose with a 3D CNN. More recent works [10, 5] apply 3D point clouds as input and can estimate very accurate hand poses.

3D data is not always available either at training or at testing. Some recent works have started to explore the use of monocular RGB data. For example, Zimmermann *et al.* [31] regress heatmaps for each hand keypoint from RGB images and then regress the 3D hand pose from these heatmaps with fully-connected layers. Mueller *et al.* [12] follow a similar approach, but obtain the final 3D hand pose by using a kinematic skeleton model to fit the probability distribution of predicted heat maps.

More recent monocular RGB-based methods leverage depth information for training [2, 19], even though testing is done exclusively with RGB images. Our proposed method also falls into this line of work. Cai *et al.* [2] propose an additional decoder to render depth maps from corresponding poses to regularize the learning of an RGB-based pose estimation system. This architecture is essentially two independent networks with a shared hand pose layer. This shared layer however cannot leverage data without pose annotations. Spurr *et al.* [19] propose a VAE-based method that learns a shared latent space for hand poses from both RGB and depth images. However, its alternating training strategy from the different modalities ignores the availability of corresponding data and leads to a slow convergence speed.

3. Methodology

The aim of cross-modal methods is to capture relationships between different modalities so that it is possible to obtain information of target modalities given observations of some other modalities. In this section, we first present the cross modal VAE (CrossVAE) [15, 19] and our extensions to handle inputs and outputs from multiple modalities (Sec. 3.1). We then introduce two latent space alignment operators strategies (Sec. 3.2) and how they can be applied for RGB-based hand pose estimation (Sec. 3.3).

3.1. Cross Modal VAE and its extension

Given data sample $\mathbf{x}$ from some input modality, the cross modal VAE aims to estimate its corresponding target value $\mathbf{y}$ in a target modality by maximizing the evidence lower bound (ELBO) via a latent variable $\mathbf{z}$.

$$\begin{aligned}\log p(\mathbf{y}) &\geq \text{ELBO}_{\text{cVAE}}(\mathbf{x};\mathbf{y};\theta,\phi) \\ &= E_{\mathbf{z}\sim q_\phi}\log p_\theta(\mathbf{y}|\mathbf{z}) - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})).\end{aligned} \tag{1}$$

Here, $D_{KL}(\cdot)$ is the Kullback-Leibler divergence. β is a hyperparameter introduced by [8] to balance latent space capacity and reconstruction accuracy. $p(\mathbf{z}) = \mathcal{N}(\mathbf{0},\mathbf{I})$ is a Gaussian prior on the latent variable $\mathbf{z}$. The variational approximation $q_\phi(\mathbf{z}|\mathbf{x})$ is an encoder from $\mathbf{x}$ to $\mathbf{z}$, and $p_\theta(\mathbf{y}|\mathbf{z})$ is a decoder or inference network from $\mathbf{z}$ to $\mathbf{y}$.

In addition to $\mathbf{x}$ and $\mathbf{y}$, we assume that there are corresponding data from N other modalities $\{\mathbf{w}_1,\ldots,\mathbf{w}_N\}$ and that these modalities are conditionally independent given latent representation $\mathbf{z}$. For clarity, we limit our derivation below to $N=1$, though the theory generalizes to higher N as well. To encode these additional modalities, we can extend the ELBO from Eq. 1 as follow:

$$\begin{aligned}\log\ p(\mathbf{y},\mathbf{w}_1) &\geq \text{ELBO}_{\text{cVAE}}(\mathbf{x},\mathbf{w}_1;\mathbf{y},\mathbf{w}_1;\phi_{\mathbf{x},\mathbf{w}_1},\theta_\mathbf{y},\theta_{\mathbf{w}_1}) \\ &= E_{\mathbf{z}\sim\phi_{\mathbf{x},\mathbf{w}_1}}\log p_{\theta_\mathbf{y}}(\mathbf{y}|\mathbf{z}) + \lambda_{\mathbf{w}_1} E_{\mathbf{z}\sim\phi_{\mathbf{x},\mathbf{w}_1}}\log p_{\theta_{\mathbf{w}_1}}(\mathbf{w}_1|\mathbf{z}) \\ &\quad -\beta D_{KL}\left(q_{\phi_{\mathbf{x},\mathbf{w}_1}}(\mathbf{z}|\mathbf{x},\mathbf{w}_1)||p(\mathbf{z})\right),\end{aligned} \tag{2}$$

where $\lambda_{\mathbf{w}_1}$ is a hyperparameter that regulates the reconstruction accuracy between $\mathbf{w}_1$ and $\mathbf{y}$. Graphical models of the original cross modal VAE and its extension to more modalities are shown in Fig 2a and Fig 2b.

We expect the $\mathbf{z}$ sampled from the variational approximation $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{w}_1)$ in Eq. 2 to be more informative than the one sampled from $q_\phi(\mathbf{z}|\mathbf{x})$ in Eq. 1, since it is conditioned on both $\mathbf{z}$ and $\mathbf{w}_1$. Furthermore, the expectation term for the decoder $p_{\theta_{\mathbf{w}_1}}$ can be regarded as a regularizer that prevents the latent space from over-fitting to $\mathbf{y}$'s modality. From here onwards, ,we define $\mathbf{z}_{\text{joint}}$ as $\mathbf{z}$ from Eq. 2.

Note that Eq. 2 assumes that corresponding data from modalities $\mathbf{x}$, $\mathbf{w}_1$ are always available. While this is a reasonable assumption for training, *i.e.* having corresponding data samples from multiple modalities, this severely limits the applicability.

Algorithm 1 Extended cross modal with one encoder.

Require: $\mathbf{x},\mathbf{y},\mathbf{w}_1,T$
Ensure: $\phi_\mathbf{x},\theta_\mathbf{y},\theta_{\mathbf{w}_1}$
1: Initialize $\phi_\mathbf{x},\theta_\mathbf{y},\theta_{\mathbf{w}_1}$
2: **for** $t=1,\ldots,T$ epochs **do**
3: Encode $\mathbf{x}$ to $q_{\phi_\mathbf{x}}(\mathbf{z}_\mathbf{x}|\mathbf{x})$
4: Decode $\mathbf{z}_\mathbf{x}$ to $p_{\theta_\mathbf{x}}(\mathbf{y}|\mathbf{z}_\mathbf{x}), p_{\theta_{\mathbf{w}_1}}(\mathbf{w}_1|\mathbf{z}_\mathbf{x})$
5: Update $\phi_\mathbf{x},\theta_\mathbf{y},\theta_{\mathbf{w}_1}$ via gradient ascent of $\text{ELBO}_{\text{cVAE}}(\mathbf{x};\mathbf{y},\mathbf{w}_1;\phi_\mathbf{x},\theta_\mathbf{y},\theta_{\mathbf{w}_1})$
6: **end for**

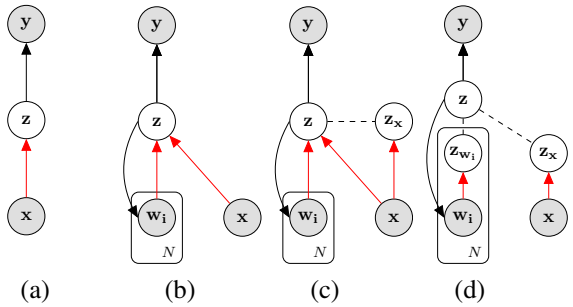

Figure 2: Graphical models. (a) Cross modal; (b) Extended cross modal; (c) Latent alignment with a KL divergence loss; (d) Latent alignment with the product of Gaussian experts. The shaded nodes represent observed variables while un-shaded nodes are latent. The red and black solid lines denote variational approximations q_ϕ or encoders, and the generative models p_θ or decoders respectively. The dashed lines denote the operation that embedding cross-modal latent spaces into a joint shared latent space; it is a KL divergence optimization for (c) and product of Gaussian experts for (d). Figure best viewed in colour.

One possibility is to simplify the encoder to take only inputs from $\mathbf{x}$, so that Eq. 2 simplifies to $\text{ELBO}_{\text{cVAE}}(\mathbf{x};\mathbf{y},\mathbf{w}_1;\phi_\mathbf{x},\theta_\mathbf{y},\theta_{\mathbf{w}_1})$. The associated algorithm is shown in Alg. 1. Note that this reduces the richness of the latent space and thereby the decoding capabilities.

3.2. Latent Space Alignment

An alternative solution is to learn $q_{\phi_{\mathbf{x},\mathbf{w}_1}(\mathbf{z}|\mathbf{x},\mathbf{w}_1)}$ and $q_{\phi_\mathbf{x}(\mathbf{z}|\mathbf{x})}$ jointly and ensure that they correspond, *i.e.* are equivalent, by aligning the two distributions together. Note that equivalence between the two distributions follows naturally from our originally assumption that $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{w}_i$ are all conditionally independent given $\mathbf{z}$. Inspired by multimodal learning work of [1], we propose joint training objectives to align the latent spaces learned from single modalities to the one learned with joint modalities to improve inference capabilities. More specifically, we would like to align $\mathbf{z}_\mathbf{x}$ (the latent representation learned only from $\mathbf{x}$), with the joint latent representation $\mathbf{z}_{\text{joint}}$ learned from both $\mathbf{x}$ and $\mathbf{w}$ so as to leverage the modalities of $\mathbf{w}$. One can also regard this as bringing together $q_{\phi_{\mathbf{x},\mathbf{w}_1}(\mathbf{z}|\mathbf{x},\mathbf{w}_1)}$ and $q_{\phi_\mathbf{x}(\mathbf{z}|\mathbf{x})}$ as close as possible.

KL divergence Loss. An intuitive way of aligning one latent space with another is to incorporate an additional loss term to reduce the divergence between $q_{\phi_{\mathbf{x},\mathbf{w}_1}(\mathbf{z}|\mathbf{x},\mathbf{w}_1)}$ and $q_{\phi_\mathbf{x}(\mathbf{z}|\mathbf{x})}$. This was first proposed by [21] for handling missing data from input modalities in multimodal setting. While we have no missing data in our cross-modal setting, we introduce a similar KL-divergence term D_{KL} with hyper-

parameter β' to align the latent spaces.

$$\begin{aligned}\mathcal{L}(\phi_{\mathbf{x},\mathbf{w}_1},\phi_{\mathbf{x}},\theta_{\mathbf{y}},\theta_{\mathbf{w}_1}) \quad (3)\\ = \text{ELBO}_{\text{cVAE}}(\mathbf{x},\mathbf{w}_1;\mathbf{y},\mathbf{w}_1;\phi_{\mathbf{x},\mathbf{w}_1},\theta_{\mathbf{y}},\theta_{\mathbf{w}_1})\\ + \text{ELBO}_{\text{cVAE}}(\mathbf{x};\mathbf{y},\mathbf{w}_1;\phi_{\mathbf{x}},\theta_{\mathbf{y}},\theta_{\mathbf{w}_1})\\ - \beta' D_{KL}\left(q_{\phi_{\mathbf{x},\mathbf{w}_1}}(\mathbf{z}_{\text{joint}}|\mathbf{x},\mathbf{w}_1)||q_{\phi_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x})\right).\end{aligned}$$

Note that the decoders $\theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$ are shared in the above ELBOs in Eq. 3. This implicitly forces $\mathbf{z}_{\text{joint}}$ and $\mathbf{z}_{\mathbf{x}}$ to be embedded to the same space (see Fig. 2c and Alg. 2).

The above formulation suffers from two major drawbacks on the encoding side. Firstly, as the number of modalities or N increases, the joint encoder $q_{\phi_{\mathbf{x},\mathbf{w}_1}}$ becomes difficult to learn. Secondly, with only the two encoders $q_{\phi_{\mathbf{x}}}$ and $q_{\phi_{\mathbf{x},\mathbf{w}_1}}$, we are not able to leverage data pairs ($\mathbf{w}_1$,$\mathbf{y}$). To overcome these weaknesses, we introduce the product of experts (PoE) as an alternative form of alignment.

Product of Gaussian Experts. It was proven in [26] that the joint posterior is proportional to the product of individual posteriors, *i.e.* $q(\mathbf{z}|\mathbf{x},\mathbf{w}_1) \propto p(\mathbf{z})q(\mathbf{z}|\mathbf{x})q(\mathbf{z}|\mathbf{w}_1)$. To that end, we can estimate the joint latent representation from unimodal latent representations. Recall that in the formulation of the VAE, both $p(\mathbf{z})$ and $q(\mathbf{z}|\cdot)$ are Gaussian; as such, we arrive at $q(\mathbf{z}|\mathbf{x},\mathbf{w}_1)$ through a simple product of Gaussian experts, $q(\mathbf{z}|\mathbf{x})$ and $q(\mathbf{z}|\mathbf{w}_1)$ [3, 26] (see model in Fig. 2d). With the help of shared decoders, we arrive at a joint latent representation through the following objective:

$$\begin{aligned}&\mathcal{L}(\phi_{\mathbf{x}},\phi_{\mathbf{w}_1},\theta_{\mathbf{y}},\theta_{\mathbf{w}_1}) = \text{ELBO}_{\text{cVAE}}(\mathbf{x};\mathbf{y},\mathbf{w}_1;\phi_{\mathbf{x}},\theta_{\mathbf{y}},\theta_{\mathbf{w}_1})\\ &+ \text{ELBO}_{\text{cVAE}}(\mathbf{w}_1;\mathbf{y},\mathbf{w}_1;\phi_{\mathbf{w}_1},\theta_{\mathbf{y}},\theta_{\mathbf{w}_1}) \quad (4)\\ &+ \text{ELBO}_{\text{cVAE}}(\mathbf{x},\mathbf{w}_1;\mathbf{y},\mathbf{w}_1;\phi_{\mathbf{x}},\phi_{\mathbf{w}_1},\theta_{\mathbf{y}},\theta_{\mathbf{w}_1})\\ &= E_{\mathbf{z}_{\mathbf{x}}\sim q_{\phi_{\mathbf{x}}}}\log p_\theta(\mathbf{y},\mathbf{w}_1|\mathbf{z}_{\mathbf{x}}) + E_{\mathbf{z}_{\mathbf{w}_1}\sim q_{\phi_{\mathbf{w}_1}}}\log p_\theta(\mathbf{y},\mathbf{w}_1|\mathbf{z}_{\mathbf{w}_1})\\ &+ E_{\mathbf{z}_{\text{joint}}\sim\text{GProd}(\mathbf{z}_{\mathbf{x}},\mathbf{z}_{\mathbf{w}_1})}\log p_\theta(\mathbf{y},\mathbf{w}_1|\mathbf{z}_{\text{joint}})\\ &- \beta(D_{KL}\left(q_\phi(\mathbf{z}_{\mathbf{x}}|\mathbf{x})||p(\mathbf{z})\right) + D_{KL}\left(q_\phi(\mathbf{z}_{\mathbf{w}_1}|\mathbf{w}_1)||p(\mathbf{z})\right)),\end{aligned}$$

where the GProd$(\cdot)$ is the product of Gaussian experts. Note in this formulation, we do not need a joint encoder $\phi_{\mathbf{x},\mathbf{w}_1}$ for $\mathbf{x}$ and $\mathbf{w}_1$ as was the case for alignment with KL divergence in Eq. 3. Instead, we use $q(\mathbf{z}|\mathbf{x})$ and $q(\mathbf{z}|\mathbf{w}_1)$ as two Gaussian experts. Suppose that $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_1,\Sigma_1)$ and $q(\mathbf{z}|\mathbf{w}_1) = \mathcal{N}(\mu_2,\Sigma_2)$. The product of two Gaussian experts is also Gaussian with mean μ and covariance Σ, where

$$\mu = (\mu_1 T_1 + \mu_2 T_2)/(T_1+T_2), \quad \text{and} \quad (5)$$

$$\sigma = 1/(T_1+T_2), \quad \text{where } T_1 = 1/\Sigma_1, T_2 = 1/\Sigma_2. \quad (6)$$

All operations in the product of Gaussian experts are element-wise. In this way, we can build a connection between $\mathbf{z}_{\text{joint}}$ and $\mathbf{z}_{\mathbf{x}}, \mathbf{z}_{\mathbf{w}_1}$, forcing them all into one shared latent space. This alignment strategy is more flexible than Alg. 2, because the encoders of different modalities can be trained individually, even from different datasets, while for Alg. 2, the joint encoder must be trained on the complete $\mathbf{x}, \mathbf{w_1}$ pairs. The learning algorithm can be found in Alg. 3.

3.3. Application Towards Hand Pose Estimation

In the context of RGB-based hand pose estimation, $\mathbf{x}$ represents RGB images and $\mathbf{y}$ 3D hand poses. Other modalities like heatmaps, depth maps, point clouds and segmentation masks can be used as $\mathbf{w}$ during training to improve the learning of the latent space and thereby leading to more accurate hand pose estimates from RGB inputs. In this paper, we use point clouds (C) and heat maps (H) as additional modalities $\mathbf{w}$ to improve the cross modal inference of RGB (R) to 3D poses (P). In the rest of paper, we use the format "A2B" to represent the estimation of target modality "B" from input modality "A" during training. For example, R2CHP represents the estimation of point clouds, heat maps and 3D poses from RGB input. Note that unless indicated otherwise, the test settings use RGB images as the source modality or input and 3D hand poses as the target modality or output.

4. Implementation Details

4.1. Data Pre-Processing and Augmentation

From the RGB image, the region containing hand is cropped from ground truth masks and resized to 256×256. The corresponding region in the depth image is converted

Algorithm 2 Latent alignment with Eq. 3.

Require: $\mathbf{x}, \mathbf{y}, \mathbf{w}_1, T$
Ensure: $\phi_{\mathbf{x}}, \phi_{\mathbf{x},\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$
1: Initialize $\phi_{\mathbf{x}}, \phi_{\mathbf{x},\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$
2: **for** $t = 1,\ldots,T$ epochs **do**
3: Encode $\mathbf{x}$ to $q_{\phi_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x})$
4: Encode $\mathbf{x}, \mathbf{w}_1$ to $q_{\phi_{\mathbf{x},\mathbf{w}_1}}(\mathbf{z}_{\text{joint}}|\mathbf{x},\mathbf{w}_1)$
5: Decode $\mathbf{z}_{\mathbf{x}}$ to $p_{\theta_{\mathbf{x}}}(\mathbf{y}|\mathbf{z}_{\mathbf{x}}), p_{\theta_{\mathbf{w}_1}}(\mathbf{w}_1|\mathbf{z}_{\mathbf{x}})$
6: Decode $\mathbf{z}_{\text{joint}}$ to $p_{\theta_{\mathbf{x}}}(\mathbf{y}|\mathbf{z}_{\text{joint}}), p_{\theta_{\mathbf{w}_1}}(\mathbf{w}_1|\mathbf{z}_{\text{joint}})$
7: Construct $D_{KL}(q_{\phi_{\mathbf{x},\mathbf{w}_1}}(\mathbf{z}_{\text{joint}}|\mathbf{x},\mathbf{w}_1)||q_{\phi_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x}))$
8: Update $\phi_{\mathbf{x}}, \phi_{\mathbf{x},\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$ via gradient ascent of Eq. 3
9: **end for**

Algorithm 3 Latent alignment with Eq. 4.

Require: $\mathbf{x}, \mathbf{y}, \mathbf{w}_1, T$
Ensure: $\phi_{\mathbf{x}}, \phi_{\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$
1: Initialize $\phi_{\mathbf{x}}, \phi_{\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$
2: **for** $t = 1,\ldots,T$ epochs **do**
3: Encode $\mathbf{x}$ to $q_{\phi_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}|\mathbf{x})$
4: Encode $\mathbf{w}_1$ to $q_{\phi_{\mathbf{w}_1}}(\mathbf{z}_{\mathbf{w}_1}|\mathbf{w}_1)$
5: Construct $\mathbf{z}_{\text{joint}} = \text{GProd}(\mathbf{z}_{\mathbf{x}}, \mathbf{z}_{\mathbf{w}_1})$
6: Decode $\mathbf{z}_{\mathbf{x}}, \mathbf{z}_{\mathbf{w}_1}, \mathbf{z}_{\text{joint}}$ to $p_{\theta_{\mathbf{x}}}(\mathbf{y}|\cdot), p_{\theta_{\mathbf{w}_1}}(\mathbf{w}_1|\cdot)$ respectively
7: Update $\phi_{\mathbf{x}}, \phi_{\mathbf{w}_1}, \theta_{\mathbf{y}}, \theta_{\mathbf{w}_1}$ via gradient ascent of Eq. 4
8: **end for**

to point clouds using the provided camera intrinsic parameters. For each training step, a different set of 256 points are randomly sampled as training input.

Viewpoint correction. After cropping the hand from the RGB image, the center of the hand in the image moves from some arbitrary coordinates to the center of the image. As such, the 3D hand pose and associated point cloud must be rotated such that the viewing angle towards the hand aligns with the optical axis. As indicated in [10], this correction is necessary to remove the many-to-one observation-pose pairings. We follow the approach given in [10]. Detailed equations on view correction can be found in the supplementary material.

Data augmentation was performed online during training. The images are scaled randomly between $[1, 1.2]$, translated $[-20, 20]$ pixels and rotated $[-\pi, \pi]$ around the camera view axis. Furthermore, the hue of the image is randomly adjusted by [-0.1, 0.1]. The point clouds are rotated randomly around the camera view axis and the 3D pose labels are also rotated accordingly.

4.2. Encoder and Decoder Modules

Our proposed method is highly flexible and can integrate many different modalities to construct a common latent space. In the current work, we learn encoders for RGB images and point clouds and decoders for 3D hand poses, point clouds and heat maps of the 2D hand key points on the RGB image. We choose to convert the 2.5D depth information as 3D point clouds instead of standard depth maps, due to its superior performance in hand pose estimation, as shown in previous works [10, 4, 6]. Heat maps are chosen as a third modality for decoding to encourage convergence of the RGB encoder, since the heat maps are closely related to activation areas on the RGB images.

For encoding RGB images, we use Resnet-18 from [7] and two additional fully connected layers to predict the mean and variance vector of the latent variable. For encoding point clouds, we employ the ResPEL network [10], which is an learning architecture that takes unordered point cloud as input. While we use same number of PEL layers as in [10], the number of hidden units are reduced by half to ease the computational load.

To decode the heatmaps, we follow the decoder architecture of the DC-GAN [18]. The loss function used for the heatmaps is the L2 loss function of pixel-wise difference between prediction and ground-truth:

$$\mathcal{L}_{\text{heat}} = \sum_{j=1}^{J} ||\hat{H}_j - H_j||, \tag{7}$$

whereas H_j is the ground-truth heatmap for the j-th hand keypoint and $\hat{H}_j$ is the prediction. For decoding point clouds, we follow the FoldingNet architecture [28] and try to reconstruct a point cloud representing the visible surface of the hand. To learn the decoder, we use two different loss terms based on the Chamfer distance and Earth Mover's distance (EMD). The Chamfer distance is the sum of the Euclidean distance between points from one set and its closest point in the other set and vice versa:

$$\mathcal{L}_{\text{Chamfer}} = \frac{1}{|P|}\sum_{p\in P}\min_{\hat{p}\in\hat{P}}||\hat{p}-p|| + \frac{1}{|\hat{P}|}\sum_{\hat{p}\in\hat{P}}\min_{p\in P}||\hat{p}-p||. \tag{8}$$

For the Earth Mover's distance, one-to-one bijective correspondences are established between two point clouds, and the Euclidean distances between them are summed:

$$\mathcal{L}_{\text{EMD}} = \min_{\phi:P\to\hat{P}}\frac{1}{|P|}\sum_{p\in P}||p-\phi(p)||, \tag{9}$$

In both Eq. 8 and 9, $\hat{P}, P \in \mathbb{R}^3$ represent the predicted point clouds and the ground truth point clouds respectively and the number of points in both clouds are 256.

The decoder for 3D pose consists of 4 fully-connected layers with 128 hidden units for each layer. To learn the pose decoder, we use an L2 loss:

$$\mathcal{L}_{\text{pose}} = ||\hat{y} - y||, \tag{10}$$

where $\hat{y}, y$ are the predicted and the ground truth hand poses describing the 3D locations of 21 keypoints.

Combining all the losses in Eq. 7-10, we obtain the following reconstruction loss function:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{pose}} + \lambda_{\text{heat}}\mathcal{L}_{\text{heat}} + \lambda_{\text{cloud}}(\mathcal{L}_{\text{Chamfer}} + \mathcal{L}_{\text{EMD}}). \tag{11}$$

The overall loss for training is the sum of reconstruction loss and its corresponding D_{KL} loss based on Eq. 2-4.

5. Experimentation

In the experiments, we set the dimensionality of latent variable $\mathbf{z}$ to 64, λ_{heat} to 0.01, λ_{cloud} to 1 for all cases and β' to 1 for Eq. 3 . Our method is implemented with Tensorflow. For learning, we use an Adam optimizer with an initial learning rate of 10^{-4} and a batch size of 32. We lower the learning rate by a factor of 10 two times after convergence. The value of β is annealed from 10^{-5} to 10^{-3}.

5.1. Datasets and evaluation metrics

Our method is evaluated on two publicly available datasets: the Rendered Hand Pose Dataset (RHD) [31] and the Stereo Hand Pose Tracking Benchmark (STB) [30].

RHD is a synthesized dataset of rendered hand images with 320×320 resolution from 20 characters performing 39

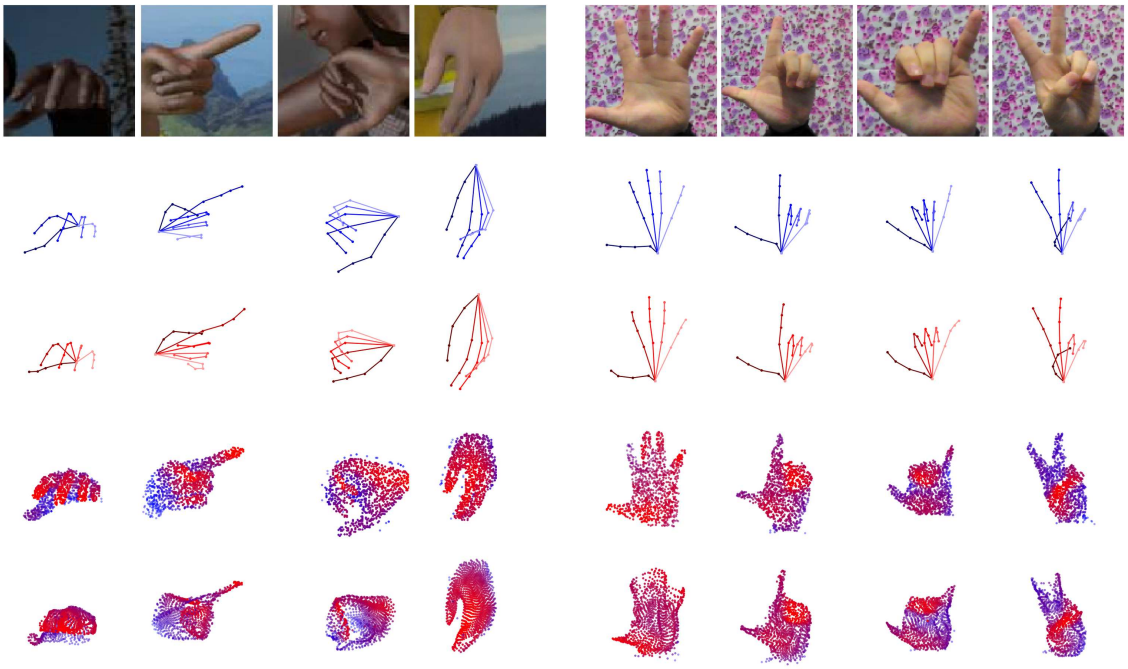

Figure 3: 3D pose estimation and point cloud reconstruction for RHD (left) and STB (right) dataset. From top to bottom: RGB images, ground-truth poses in blue, estimated poses from $\mathbf{z}_{rgb}$ in red, ground-truth point clouds, reconstructed point clouds from $\mathbf{z}_{rgb}$. The color for point clouds decodes the depth information, closer points are more red and further points are more blue. Note that the ground-truth point clouds are not used for inference, it is shown here only for comparison purpose.

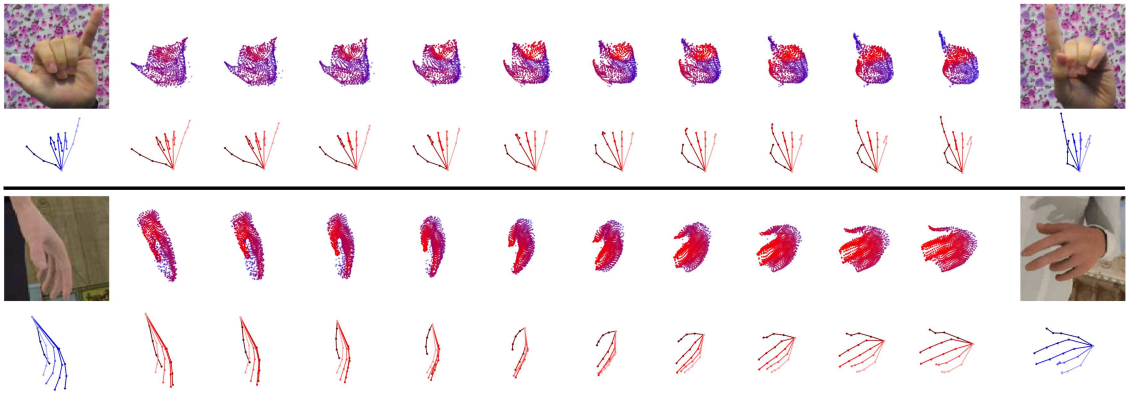

Figure 4: Latent space interpolation. Two examples of reconstructing point clouds and hand poses from the latent space. The most left and most right column are RGB images and their corresponding ground-truth poses. Other columns are generated point clouds and poses when interpolating linearly on the latent space.

actions. It is composed of 41238 samples for training and 2728 samples for testing. For each RGB image, a corresponding depth map, segmentation mask, and 3D hand pose are provided. The dataset is highly challenging because of the diverse visual scenery, illumination, and noise.

STB contains videos of a single person's left hand in front of six different real-world backgrounds. The dataset provides stereo images, color-depth pairs with 640 $\times$ 480 resolution and 3D hand pose annotations. Each of the 12 sequences in the dataset contains 1500 frames. To make the 3D pose annotations consistent for RHD, we follow [31, 2] and modify the palm joint in STB to the wrist point. Similar to [31, 2, 19, 27], we use 10 sequences for training and the other 2 for testing.

To evaluate the accuracy of the estimated hand poses, we use the common metrics mean end-point-error (EPE) and area under the curve (AUC) on the percentage of correct keypoints (PCK) curve. EPE is measured as the average Eu-

Strategy	Encoder	Decoder	Mean EPE [mm]
S1 (Eq. 1)	R	P	16.61
S2 (Alg. 1)	R	H+P	16.10
	R	C+P	15.91
	R	C+H+P	15.49
S3 (Alg. 2)	R+C	C+H+P	14.93
S4 (Alg. 3)	R+C	C+H+P	**13.14**

Table 1: Comparison of different training strategies on the RHD dataset. The mean EPE values are obtained from monocular RGB images. (R: RGB, C: point cloud, P: pose, H: heatmap). Poses estimated from monocular RGB images can be improved by increasing number of different encoders and decoders during training.

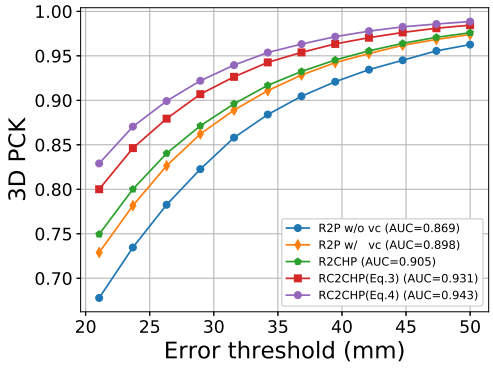

Figure 5: Comparisons of 3D PCK results of our different strategies on RHD dataset. The abbreviations can be found in Sec. 3.3 and "vc" stands for "view correction"

clidean distance between predicted and ground-truth hand joints, whereas AUC represents the percentage of predicted keypoints that fall within certain error thresholds compared with ground-truth poses. To compare with the state-of-the-art methods in a fair way, we follow the similar condition used in [19, 9, 2, 27] to assume that the global hand scale and the hand root position are known in the experimental evaluations, where we set the middle finger's base position as the root of the hand.

5.2. Qualitative results

Using the flexible design of our method, we train the networks exploiting all the available modalities and test using only limited modalities. In Fig. 3, we show some qualitative examples of poses and point clouds decoded from the $\mathbf{z}_{\text{rgb}}$. The 3D poses and point clouds can be successfully reconstructed from the same latent variable $\mathbf{z}$. The reconstructed point clouds' surfaces are smoother than the original inputs, since the inputs are sub-sampled from raw sensor data, while the reconstructed point clouds hold some structured properties from the FoldingNet decoder.

We also evaluate the ability of our model to synthesize hand poses and point clouds. From two RGB images of the hand, we estimate the corresponding latent variables $\mathbf{z}_{1,2}$ and then sample points by linearly interpolating between the two. 3D hand pose and point cloud reconstructions of the interpolated points via our learned decoders are shown in Fig. 4. We observe that the learned latent space reconstructs a smooth and realistic transition between different poses, with changes in both global rotations and local finger configurations.

5.3. RGB 3D Hand Pose Estimation

Note that even though our network is trained with multiple modalities, the results provided here are based only in monocular RGB inputs.

Training Strategy. We first compare different training strategies (S) in Table 1: S1. Baseline method to only use RGB-pose pairs for training. S2. Training with extended decoders, where the latent variables $\mathbf{z}_{\text{rgb}}$ reconstruct more modalities (heatmaps and point clouds) besides poses. S3. Training with an additional encoder for point clouds, where the different latent variables are aligned as per Alg. 2. S4. The alignment method in S3 is changed to the product of Gaussian experts (Alg. 3). More comparison results with AUC metric are shown in Fig. 5

Comparing S1 to the other strategies, we observe that the baseline performance can be improved by training with increasing number of additional encoders or decoders. Comparing S4 to S3, the alignment with the Gaussian product outperforms the intuitive KL-divergence alignment method by capturing a better joint posterior of different input modalities.

Furthermore, we emphasize the necessity of viewpoint correction (Sec. 4.1). We applied both view corrected and uncorrected data for training the baseline strategy "R2P" (S1). The difference can be seen from Fig. 5, where the view corrected data clearly improves the AUC metric.

	Method	RHD	STB
VAE-based	Spurr *et al.* [19]	19.73	8.56
	Yang *et al.* [27]	19.95	8.66
	Ours	**13.14**	**7.05**
Others	Z&B [31]	30.42	8.68
	Iqbal *et al.* [9]	13.41	\

Table 2: Comparison to state-of-the-art on the RHD and STB with mean EPE [mm]. Ours refers to S4 in Table 1 (RC2CHP).

Comparison to state-of-the-art. In Table 2, we compare the EPE of our method with VAE-based methods [19, 27] which are most related to our method as well as other

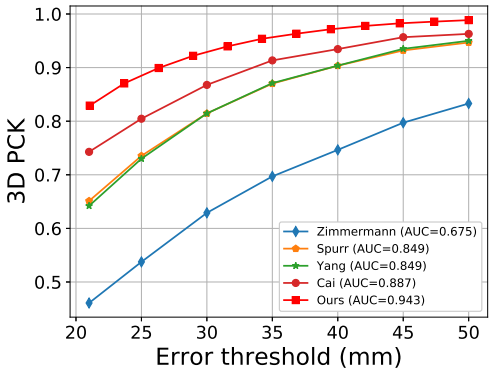

Figure 6: AUC: Comparison to state-of-the-art methods on the RHD dataset. Ours refers to S4 in Table 1 (RC2CHP).

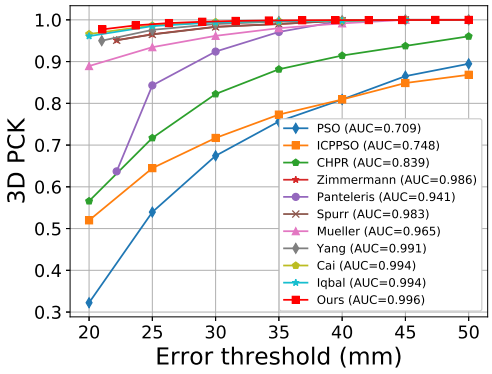

Figure 7: AUC: Comparison to state-of-the-art methods on the STB dataset. Ours refers to S4 in Table 1 (RC2CHP).

state-of-the-art [31, 9]. On both datasets, our proposed method achieves the best results, including an impressive 1.61mm or 19% improvement on the STB dataset.

We also compare the PCK curve of our approach with other state-of-the-art methods [19, 27, 31, 9, 12, 16] in Fig. 6 and Fig. 7. For both datasets, our method achieves the highest AUC value on the 3D PCK. We marginally outperform the state-of-the-art [9, 2] on the STB dataset, whereas on the RHD dataset, we surpass all reported methods to date [31, 27, 2, 19] with a significant margin. We note, however, that the STB dataset contains much less variation in hand poses and backgrounds than the RHD dataset and that performance by state-of-the-art methods on STB has become saturated. As such, there is little room for improvement on STB, whereas the benefits of our method is more visible on the RHD dataset.

Weakly-supervised learning. Thanks to flexibility of the proposed method, (surface) point clouds can be also used as "weak" labels for unlabelled data to aid the training process. We tested our method under a weakly-supervised setting on the RHD dataset, where we sample the first m% samples as labelled data (including RGB, point clouds and 3D poses) and the rest as unlabelled data (including RGB, point clouds) by discarding 3D pose labels. We compare the supervised setting with the weakly-supervised setting for the "RC2CHP" networks (S4 in Table 1). In the supervised training setting, we train the networks with only m% samples, In the weakly-supervised setting, besides fully supervised training on m% data, we also train the "RC2C" sub-parts with the rest (100-m)% samples simultaneously. The percentage of labelled data is varied from 5% to 100% to compare the mean EPE between supervised and weakly-supervised settings. From Fig. 8 we can see that our method makes full usage of additional unlabelled information, where the improvement is up to 6%.

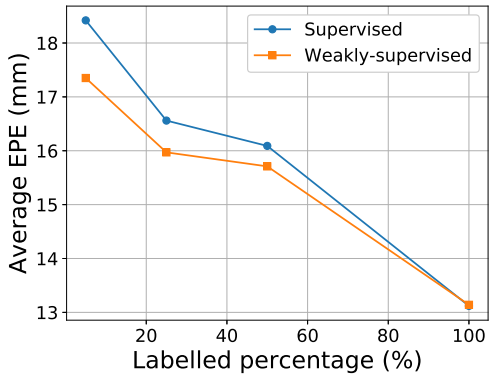

Figure 8: Mean EPE of our model on the weakly-supervised setting. Our method makes full use of unlabelled data, as the weakly-supervised setting performs almost as well as the supervised one.

6. Conclusion

In this paper, we formulate RGB-based hand pose estimation as a multimodal learning and cross-modal inference problem. We derive different objectives for three hand modalities, and show different ways of aligning their associated latent spaces with a joint one. Our experiments show that the proposed method can exploit different modalities as prior knowledge to improve the performance of RGB-based hand pose estimation as well as leverage weakly labelled data. Experiments on two publicly available datasets demonstrate that our approach outperform previous state-of-the-art methods. Moreover, the model size and runtime of our architecture is kept the same as other VAE-based hand estimation methods at test time.

Acknowledgments Research in this paper was partly supported by the Singapore Ministry of Education Academic Research Fund Tier 1. We thank the Helmholtz Association for support. We also gratefully acknowledge NVIDIA's donation of a Titan X Pascal GPU.

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. 3

[2] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, 2018. 1, 2, 6, 7, 8

[3] Yanshuai Cao and David J Fleet. Generalized product of experts for automatic and principled fusion of gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014. 4

[4] Xinghao Chen, Guijin Wang, Cairong Zhang, Tae-Kyun Kim, and Xiangyang Ji. Shpr-net: Deep semantic hand pose regression from point clouds. *IEEE Access*, 6:43425–43439, 2018. 5

[5] Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *CVPR*, 2018. 2

[6] Liuhao Ge, Zhou Ren, and Junsong Yuan. Point-to-point regression pointnet for 3d hand pose estimation. In *ECCV*, pages 475–491, 2018. 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[8] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 3

[9] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, 2018. 7, 8

[10] Shile Li and Dongheui Lee. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In *CVPR*, 2019. 1, 2, 5

[11] 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*, pages 5079–5088, 2018. 2

[12] 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, 8

[13] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. In *WACV*, 2015. 2

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

[15] Gaurav Pandey and Ambedkar Dukkipati. Variational methods for conditional multimodal deep learning. In *IJCNN*, 2017. 2

[16] 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. 8

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

[18] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 5

[19] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 2, 6, 7, 8

[20] James S Supancic, Grégory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan. Depth-based hand pose estimation: data, methods, and challenges. In *ICCV*, 2015. 1

[21] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016. 2, 3

[22] Anastasia Tkach, Andrea Tagliasacchi, Edoardo Remelli, Mark Pauly, and Andrew Fitzgibbon. Online generative model personalization for hand tracking. *ACM Transactions on Graphics (TOG)*, 36(6):243, 2017. 2

[23] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (TOG)*, 33(5):169, 2014. 2

[24] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*, 2017. 2

[25] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Crossing nets: Combining GANs and VAEs with a shared latent space for hand pose estimation. In *CVPR*, 2017. 1

[26] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *NIPS*, 2018. 2, 4

[27] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *CVPR*, 2019. 1, 2, 6, 7, 8

[28] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, 2018. 1, 5

[29] Shanxin Yuan, Qi Ye, Guillermo Garcia-Hernando, and Tae-Kyun Kim. The 2017 hands in the million challenge on 3d hand pose estimation. *arXiv:1707.02237*, 2017. 2

[30] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *ICIP*, 2017. 5

[31] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017. 1, 2, 5, 6, 7, 8