3D Hand Pose Estimation with Disentangled Cross-Modal Latent Space

Jiajun Gu[1] Zhiyong Wang[1] Wanli Ouyang[1] Weichen Zhang[1] Jiafeng Li[2] Li Zhuo[2]
[1] The University of Sydney [2] Beijing University of Technology
jigu6612@uni.sydney.edu.au {zhiyong.wang, wanli.ouyang, weichen.zhang}@sydney.edu.au
{lijiafeng, zhuoli}@bjut.edu.cn

Abstract

Estimating 3D hand pose from a single RGB image is a challenging task because of its ill-posed nature (i.e., depth ambiguity). Recently, various generative approaches have been proposed to predict the 3D joints of an RGB hand image by learning a unified latent space between two modalities (i.e., RGB image and 3D joints). However, projecting multi-modal data (i.e., RGB images and 3D joints) into a unified latent space is difficult as the modality-specific features usually interfere the learning of the optimal latent space. Hence in this paper, we propose to disentangle the latent space into two sub-latent spaces: modality-specific latent space and pose-specific latent space for 3D hand pose estimation. Our proposed method, namely Disentangled Cross-Modal Latent Space (DCMLS), consists of two variational autoencoder networks and auxiliary components which connect the two VAEs to align underlying hand poses and transfer modality-specific context from RGB to 3D. For the hand pose latent space, we align it with the two modalities by using a cross-modal discriminator with an adversarial learning strategy. For the context latent space, we learn a context translator to gain access to the cross-modal context. Experimental results on two widely used public benchmark datasets RHD and STB demonstrate that our proposed DCMLS method is able to clearly outperform the state-of-the-art ones on single image based 3D hand pose estimation.

1. Introduction

Hand pose estimation [30, 5, 36, 24, 12, 8, 31, 28] has been widely studied for various applications in augmented reality, virtual reality and human machine interaction systems. Estimating 3D hand pose from a single image remains a challenge because hands are highly deformable. The challenging aspects such as complex hand gestures, different viewpoints and hand articulations lead to further difficulties in inferring precise joint positions in 3D space [2, 28, 9]. Hand-crafted anatomical models were deployed to address these issues. Many single depth image based 3D hand pose estimation methods [31, 30, 29, 23, 33, 32] have been proposed with promising performance recently, as depth information makes it easier to deal with cluttered background. Due to the easy access to RGB data through conventional cameras, 3D hand pose estimation task using RGB images [43, 27, 22, 6, 15, 25] has attracted lots of attentions. RGB-based methods attempt to recover 3D hand poses directly from RGB images. Nevertheless, many additional issues are exposed such as context distractions (*e.g.*, background and lighting configurations) and hand appearance variations. Hence RGB-based 3D hand pose estimation still remains an ill-posed problem due to the depth ambiguities.

A recent work by Spurr *et al*. [27] extends the idea of a depth based approach from Wan *et al*. [33], which learns a latent space of 3D hand poses, and apply it to RGB based pose estimation task by learning a single cross-modal latent space between RGB modality to 3D hand pose modality. As a by-product, it jointly learns a mapping from the input modality to the output modality. However, when performing cross-modal generation as in [27], the single latent space learnt from multiple modalities usually contains the issues of balancing the shared representation and modality-specific representation. In addition, the modality-specific features usually interfere the learning of the shared latent space. Modality context, for example captures background details for 2D RGB images and 3D camera intrinsic for 3D hand pose domain and those should not be encapsulated in a shared representation space, while these modality specific representation are essential and uniquely characterizes each modalities.

However, directly capturing these representation are not feasible since there is no clear boundary or a representation set which could distinguish one modality from the other, hence the alternative is to approximate them by exploiting single modality self-reconstruction when having a shared modality-shared representation set. To this end, we propose to address these two issues by disentangling the shared latent space and modality-specific latent space, and improve each disentangled latent space by using different modules, respectively.

In this paper, we introduce a novel 3D hand pose estimation method, namely Disentangled Cross-Modal Latent Space (DCMLS), which learns better latent space by hand pose representation disentanglement. Our model consists of two variational autoencodesr (VAEs) [17] to extract the latent representation of the two modalities. For each modality, we decompose the latent representation into two parts, including (1) shared hand pose representation and (2) modality-specific context representation. To disentangle and align the shared hand pose representation of the two modalities, we train a hand representation discriminator to learn and update the encoders by using the adversarial learning strategy [11]. Meanwhile, for modality-specific context representation, apart from learning the individual modality-specific representation by self-reconstruction, we further learn a modality translator to map the cross-modal context representation from one modality (*i.e.*, RGB) to the other (*i.e.*, 3D). After recomposing the context representation with the hand pose representation, we acquire better representation from both modality-specific and modality-shared aspects, and as a result obtain improved performance of estimating 3D hand joints.

Overall, the key contributions of our work can be summarised as follows:

- To the best of our knowledge, this is the first study on 3D hand pose estimation which disentangles the hand pose and context representation from different modality latent spaces without explicit supervision of the disentangling factors.
- We propose a novel dual VAE structured network with a cross-modal hand pose discriminator and a cross-modal context translator to improve different modality latent spaces. The hand pose discriminator aligns the shared hand pose representations across different modalities. The context translator maps the context representation from one modality image (*e.g.*, RGB or 2D) to the other (*e.g.*, 3D pose).
- We conduct comprehensive qualitative and quantitative experiments to demonstrate the superiority of our proposed DCMLS model on two public RGB datasets (i.e., RHD and STB) against the state-of-the-art ones.

2. Related Work

In this section, we review the works relevant from our topics, 3D hand pose estimation, modality transfer and disentangled representation.

2.1. 3D Hand Pose Estimation

Estimating 3D hand pose has been researched intensively in the past years. Traditional model-based approaches [2, 9, 28] focus on converting the anatomical information to tackle the hand deformation and occlusion issues. Depth based approaches [4, 34, 21, 33, 10, 26, 1] have been well studied where most methods have achieved precise predictions [38] in recent years. For RGB-based task, many works utilise the deep Convolutional Neural Networks (CNN) to extract representation from the input RGB image. Based on the model concept, these methods are sub-categorized into mainly discriminative and generative approaches.

The discriminative methods normally detect 2D heatmaps and then predict the overall 3D hand poses or regress the depth of 2D keypoints. Recently, Zimmermann *et al*. [43] proposed the first deep learning approach that decomposed the 3D hand pose estimation task into three steps. Inspired by [35], they first learned a hand mask network in order to localize the hand region image, then employed an encoder-decoder structured network to predict 2D heatmaps of hand joints, and finally predicted a canonical pose and orientation transformation matrix for 3D joints prediction. In addition, in order to learn better hand prior for segmentation and 2D to 3D lifting task, many methods [6, 22, 4] have been proposed to use additional synthesized RGB datasets with computer generated 3D annotations to augment training. The task is then to focus more on reducing the domain discrepancy between the real and synthetic data from different datasets. Cai *et al*. [6] reduced the discrepancy by introducing a network that predicts 2D depth images as additional regularization for weak supervision. Mueller *et al*. [22] applied the image-to-image translation techniques to transfer synthesized hand images closer to the real domain and additionally added a SilNet module to enable additional supervision from ground-truth hand masks. The work in [4] however provided an alternative, which manipulated the viewpoints and shapes of 3D skeletons (hand joints) and generated realistic synthesized depth images from these augmented 3D hand poses.

Meanwhile, the generative methods normally learn a generalized hand model by learning a latent space of hand representation across various modalities. Previously, Wan *et al*. [33] proposed to find potential mapping between the latent spaces of hand poses and depth images. They proposed a dual-generative network to learn the shared latent space for pose configurations and depth images. While a Variational Autoencoder was able to learn the latent distribution of 3D hand representation, they adopted the Generative Adversarial Network [11] to synthesize depth images from latent representation of the synthetic 3D labels. Inspired by [33], Spurr *et al*. [27] suggested to generate a statistic hand model by learning a unified latent space between any two modalities of the hand, such as RGB images and 3D joints. They proposed a cross-modal VAE framework, where encoders extracted modal-specific representation and decoders generated output in either modality from a sampled latent representation of either input modality. However, we argue that each modality input potentially encodes both hand pose representation, which is shared across

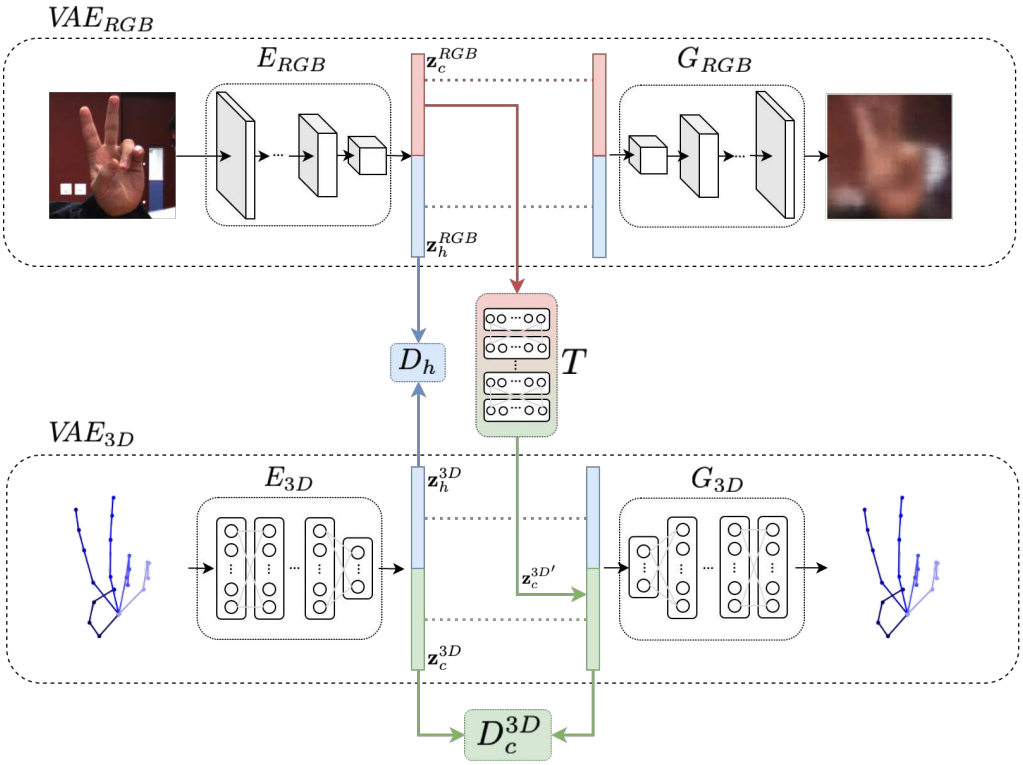

Figure 1. Overview of our proposed DCMLS model. Our model consist of two parallel VAEs (VAE_{RGB} for RGB modality and VAE_{3D} for 3D modality. The VAE of each modality consists of an encoder E and a decoder G. For learning the latent space in each modality, the embedded space of each modality z^m is disentangled into hand pose space z_h^m and modality context space z_c^m. An extra discriminator D_h connects these two VAEs by aligning the hand pose space. Lastly, we add a latent space translator T with an adversarial discriminator D_c^{3D} to transfer from one modality context to the other.

modalities, and modality specific representation, which include the context information (*e.g.*, viewpoints and backgrounds). Mapping both representations to a single cross-modal latent will result in information loss from both representations. Hence, we believe that the encoded latent representation from either modality can be disentangled into two parts to preserve both types of information, and specific modules can be used for learning each latent space effectively.

2.2. Modality transfer

Our approach is also related to modality transfer. Modality transfer techniques [16, 42, 19, 18] have been intensively utilized in the field of image-to-image translation, due to the emergence of the Generative and Adversarial Network (GAN) [11]. Generally, GAN learns the data distribution from a Gaussian noise and generates new samples by using a discriminator with the adversarial learning strategy. Based on this adversarial learning strategy, different approaches [20, 16, 42, 40, 41] which transfer the information between different modalities have been proposed. Some approaches utilising cross-modal networks [3, 14] to combine the information in different data modalities have also been proposed recently for various image generation tasks. Therefore, we propose to learn 3D hand pose configurations by utilizing the cross-modal transfer technique to map RGB modality to 3D modality. However, different from existing methods, we disentangle the latent context factors of different modalities when performing modality transfer, which is able to provide effective task specific modality transferring (i.e., 3D hand pose estimation).

2.3. Disentangled representation

In addition, in order to maximize the controlling factors for diverse and variational generation, disentangled representation has been explored to decompose these factors. While traditional GANs are unable to provide interpretable latent representation and hence unable to semantically control the generated output, Chen *et al*. [7] proposed an InfoGAN to disentangle meaningful representation from a unified latent space, such as hand writing styles for hand written numbers. Recently, a disentangled CrossVAE structure [37] was proposed to separate a number of human interpretable latent factors for both image synthesis and pose estimation task. However, instead of supervising each latent factor, we exploit the cross modal data and propose a

self-supervised approach by aligning the independent VAEs with a discriminator to disentangle the hand pose space.

3. Methodology

As shown in Figure 1, our proposed DCMLS method consist of four parts, (1) learning the single-modal latent space by utilizing a variational autoencoder (VAE) based self-reconstruction for each of the modalities (*i.e.*, 2D, 3D or RGB); (2) disentangling the latent space of each modality into two parts: modality-specific context latent space and modality-shared hand pose latent space; (3) preserving the hand pose consistency by aligning the shared hand pose representation from the two modalities with a discriminator; and (4) translating the rest of the modality specific representation by context translator from RGB modality into 3D, in order to recompose the disentangled latent space from different modalities and generate the prediction of 3D hand joints. The total objective function and overall training procedure of our DCMLS approach are introduced in Section 3.6.

3.1. Single-modal latent space learning

In general, to learn the corresponding latent space and the feature representation of a single data modality, the variational autoencoder (VAE) [17] can be effectively used. Learning a VAE from a single data modality is based on the objectives defined in [17], which is derived from maximizing the log probability (*i.e.*, $\log p(\mathbf{x})$) of the generated data samples. The main objective of learning the latent space of each modality is to maximize the reconstructed log-probability meanwhile minimizing the distance between the distribution of the latent representation and a Gaussian distribution. In the 3D hand pose estimation task, we define a set of data modalities as $M = \{\text{2D}, \text{3D}, \text{RGB}\}$. In our DCMLS model, we train a general variational autoencoder (VAE) model for each modality $m \in M$, which individually learns its corresponding latent space and feature representation. The VAE of each modality m takes their individual modality data $\mathbf{x}^m$ as the input. The basic loss function for each VAE of modality m can be defined as,

$$\mathcal{L}^m_{VAE} = \|(\mathbf{x}^m - \hat{\mathbf{x}}^m)\|^2 + \mathcal{L}_{KL}(E(\mathbf{z}^m|\mathbf{x}^m)\|p(\mathbf{z})), \quad (1)$$

where for each modality m, $\hat{\mathbf{x}}^m = G(z^m)$ is the corresponding reconstructed modality data, latent vector z is sampled from a distribution $\mathbf{z}^m(\mu, \sigma)$ by $z^m \sim (\mu, \sigma)$. E and G are respectively the encoder and decoder, $\mathbf{z}^m$ is the corresponding latent vector, $\|\cdot\|^2$ is the l_2 loss and $\mathcal{L}_{KL}$ is the KL divergence loss between the latent space distribution and the Gaussian prior distribution $p(\mathbf{z})$.

3.2. Disentangled latent space learning

In our approach, we propose to learn context-independent hand pose latent space, at the same time keep the modality context for reconstruction fidelity. As the result, in each modality m of our DCMLS model, we use a similar but different approach to disentangle two defined latent factors, which consist of the modality-specific context and the modality-shared hand pose. Hence for a modality m, we have: $\mathbf{z}^m$: $(\mathbf{z}^m_h, \mathbf{z}^m_c)$. To align with the original VAE lower bound, in each modality m, we stochastically sample two latent vectors z^m_c and z^m_h from the context and hand pose latent distributions, denoted as $\mathbf{z}^m_c$ and $\mathbf{z}^m_h$.

$$\hat{z}^m = z^m_h \oplus z^m_c, \quad (2)$$

where $\oplus$ is vector concatenation. The two decomposed latent distributions of both modality are then individually aligned with a Gaussian distribution by minimizing their KL divergence. Hence, the loss function $\mathcal{L}^m_{DL}$ of learning the disentangled latent space of each modality m is then defined by:

$$\begin{aligned}\mathcal{L}^m_{D-VAE} = \|(\mathbf{x}^m - \hat{\mathbf{x}}^m)\|_2 + \mathcal{L}_{KL}(E(\mathbf{z}^m_h|\mathbf{x}^m)\|p(\mathbf{z})) \\ + \mathcal{L}_{KL}(G(\mathbf{z}^m_c|\mathbf{x}^m)\|p(\mathbf{z})),\end{aligned} \quad (3)$$

where m stands for a specific input modality (*e.g.*, RGB, 2D or 3D) and $\mathcal{L}_{KL}$ is the same KL divergence loss in Eq. (2).

In this way, the latent space of each modality is then disentangled into two different streams, which can be both trained independently using the two modules described in the following two subsections.

3.3. Cross-modal hand pose alignment

While different modality features are captured by different modality-specific encoders, we consider to learn the shared factor from different modalities, which corresponds to the disentangled latent hand pose space. The concept of adding an auxiliary discriminator is to connect the two parallel VAEs and aligns the latent hand pose spaces. Hence, in our DCMLS model, we add a new discriminator D_h and use the adversarial learning strategy proposed in [11] to learn such embedding space shared by different modalities. The objective of aligning the hand pose specific latent space from different modalities can be achieved by optimizing (1) the loss $\mathcal{L}_{D_h}$ of training the discriminator D_h, which distinguishes the extracted representation from the two modalities x (*i.e.*, RGB, 2D) and y (*i.e.*, 3D) with

$$\begin{aligned}\mathcal{L}_{D_h} = \frac{1}{2}\mathcal{L}_{BCE}(D_h(E^{\mathbf{x}}(\mathbf{z}^{\mathbf{x}}_h|\mathbf{x})), 1) \\ + \frac{1}{2}\mathcal{L}_{BCE}(D_h(E^{\mathbf{y}}(\mathbf{z}^{\mathbf{y}}_h|\mathbf{y})), 0),\end{aligned} \quad (4)$$

and (2) the loss $\mathcal{L}_{E^{x,y}}$ of training the encoders p^x and p^y of the two modalities, which are exploited by the inversely-labelled feedback from the discriminator D_h as follows:

$$\begin{aligned}\mathcal{L}_{E^{\mathbf{x},\mathbf{y}}} = \frac{1}{2}\mathcal{L}_{BCE}(D_h(E^{\mathbf{x}}(\mathbf{z}^{\mathbf{x}}_h|\mathbf{x})), 0) \\ + \frac{1}{2}\mathcal{L}_{BCE}(D_h(E^{\mathbf{y}}(\mathbf{z}^{\mathbf{y}}_h|\mathbf{y})), 1),\end{aligned} \quad (5)$$

where $\mathcal{L}_{BCE}$ is the common binary cross entropy loss. In this way, the extracted hand pose features of one modality can be learnt to align with that of the other modality.

3.4. Cross-modal context transfer

The latent context spaces (i.e., the modality specific latent space) in different modalities are unique and quite distinct. To transfer the latent context space across modalities, the goal is to find a constrained one-to-one mapping function $F(F:\mathbf{z}_c^{\mathbf{x}} \rightarrow \mathbf{z}_c^{\mathbf{y}})$, so that the mapping function can generate latent context space $\mathbf{z}_c^{\mathbf{y}}$ of the target modality $\mathbf{y}$ (*i.e.*, 3D), given the latent space $\mathbf{z}_c^{\mathbf{x}}$ of the input modality $\mathbf{x}$ (*i.e.*, RGB or 2D). Due to the success of conditional image-to-image translation in [16], we adopt their concept to transfer the latent space across different modalities. We devise a modality specific latent space translator $T(\cdot)$ with fully connected layers as the latent space mapping function. In addition, we introduce a modality representation discriminator D_c similar as in [16] to distinguish the modality specific latent space from the original or generated modalities. Likewise, we exploit the inversely-labelled feedback of the discriminator to train the modality specific latent space translator T to transfer the latent context space from input modality $\mathbf{x}$ to target modality $\mathbf{y}$. The loss function of the overall cross-modal latent context space transfer can be expressed as follows:

$$\begin{aligned}\mathcal{L}_{D_c} = &\frac{1}{2}\mathcal{L}_{BCE}(D_c(T(E^{\mathbf{x}}(\mathbf{z}_c^{\mathbf{x}}|\mathbf{x})),0) \\ &+\frac{1}{2}\mathcal{L}_{BCE}(D_c(E^{\mathbf{y}}(\mathbf{z}_c^{\mathbf{y}}|\mathbf{y})),1),\end{aligned} \tag{6}$$

$$\begin{aligned}\mathcal{L}_T = &\|(T(E^{\mathbf{x}}(\mathbf{z}_c^{\mathbf{x}}|\mathbf{x})), E^{\mathbf{y}}(\mathbf{z}_c^{\mathbf{y}}|\mathbf{y}))\|^1 \\ &+\frac{1}{2}\mathcal{L}_{BCE}(D_h(E^{\mathbf{x}}(\mathbf{z}_c^{\mathbf{x}}|\mathbf{x})),1) \\ &+\frac{1}{2}\mathcal{L}_{BCE}(D_c(E^{\mathbf{y}}(\mathbf{z}_c^{\mathbf{y}}|\mathbf{y})),0).\end{aligned} \tag{7}$$

where $\mathcal{L}_{D_c}$ is the loss for discriminating a real and translated context space, and $\mathcal{L}_T$ is the adversarial loss for the context translator T.

3.5. Hand Pose Estimation

As we have the translator from RGB to 3D in the latent context space, we construct the 3D latent variable by concatenating the latent variable sampled from hand pose space $\mathbf{z}_h^x$ and translated context space $\hat{\mathbf{z}}_c^y$. Then, 3D data can be generated using the decoder $G^{\mathbf{y}}$ from the 3D modality by decoding the new 3D latent variables. Hence we can join the components and train our model with an end-to-end loss function:

$$\begin{aligned}\mathcal{L}_{HPE} = &\|(\mathbf{y} - G^{\mathbf{y}}(z^{\mathbf{y}}))\|_2 + \mathcal{L}_{KL}(E^{\mathbf{x}}(\mathbf{z}_h^{\mathbf{x}}|\mathbf{x})\|p(\mathbf{z})) \\ &+\mathcal{L}_{KL}(T(E^{\mathbf{y}}(\mathbf{z}_c^{\mathbf{x}})|\mathbf{y}))\|p(\mathbf{z})).\end{aligned} \tag{8}$$

3.6. Training

Overall, the total objective of our DCMLS approach combines the aforementioned loss functions in the previous subsections, which can be written as follows,

$$\begin{aligned}\min_{T,E^{\mathbf{x},\mathbf{y}},G^{\mathbf{x},\mathbf{y}}} \max_{D_h,D_c} &\mathcal{L}_{D-VAE}^{\mathbf{x}} + \mathcal{L}_{D-VAE}^{\mathbf{y}} + \mathcal{L}_{D_h} + \mathcal{L}_{D_c} \\ &\mathcal{L}_{E^{\mathbf{x},\mathbf{y}}} + \mathcal{L}_T + \mathcal{L}_{HPE},\end{aligned} \tag{9}$$

where $\mathcal{L}_{D-VAE}^{\mathbf{x}}$ and $\mathcal{L}_{D-VAE}^{\mathbf{y}}$ are the two losses of the disentangled latent space learning as in Eq. (3) for input modality $\mathbf{x}$ (*i.e.*, RGB or 2D) and target modality $\mathbf{y}$ (*i.e.*, 3D), respectively.

The training algorithm of our DCMLS approach is described in Algorithm 1. For each sample $\mathbf{x}$ from modality set X, we train the self-reconstruction pipeline with the disentangled latent space learning according to our VAE objective funciton in Eq. (3). Note that such reconstruction pipeline does not require extra labels for supervision. With paired sample x and y, we extract the latent hand features from each sample, and train the hand representation discriminator according to Eq. (4). Meanwhile, the discriminator D_H guides the encoders to extract indistinguishable hand representation from the two modalities by minimizing loss term defined in Eq. (5). In order to access the cross-modal latent space during test, we train the modality representation translator module consisting of T and D_c by using Eqs. (6 and 7). Eventually, our model combines the sampled hand representation and the translated modality representation to generate cross-modal outputs.

Algorithm 1 Disentangled Cross-Modal Latent Space.

Initialize: $\varepsilon, \theta_{E^X}, \theta_{E^Y}, \theta_{G^X}, \theta_{G^Y}, \theta_{D_h}, \theta_T, \theta_{D_c}$
for $e \leftarrow 1...\varepsilon$ epochs **do**
  for $(\mathbf{x}, \mathbf{y})$ in T **do**
    Train self-reconstruction of modality x, y by using Eq. (3).
    Train discriminator with hand representation encoded from both modalities $\mathbf{x}, \mathbf{y}$ by using Eq. (4)
    Train encoders with discriminated feedback by using Eq. (5)
    Train context translator T by using Eqs. (6, 7)
    Train cross-modal generation with transferred context $\mathbf{x}, \mathbf{y}$ by using Eq. (8)
  end for
end for

4. Experimental Results and Discussions

4.1. Datasets

We conduct experiments on two widely used public benchmarks: Stereo Hand Pose Tracking Benchmark (STB)

[39] and Rendered Hand Pose Dataset (RHD) [43].

STB is a large real dataset containing 18k frames captured with a multi-view binocular camera setup. Each frame is at the resolution of 640×480, and was annotated with the 3D joint positions of palm and fingers. With the camera intrinsic configurations provided, 2D joint positions which are aligned to the hand images can also be obtained. Although hand images were only collected from the same subject, they managed to introduce variations in terms of illumination and background conditions. To evaluate the 3D pose estimation accuracy of our model, we use the 15k/3k training/test split as in [43].

RHD is a synthetic dataset of rendered hand images containing 42k training and 2.7k testing rendered hand images with resolution of 320×320. In this dataset, hand images were taken from 20 different subjects with a total number of 39 actions. Each image was annotated with 2D and 3D joint positions. Corresponding hand mask and depth images are also available. We only used the images and their corresponding 3D labels to train our model.

4.2. Experimental settings

Our model is implemented with Pytorch[1] framework. Similar as in our baseline [27], for RGB images, we utilise random-initialized ResNet-18 [13] network as our image encoder, and use several transposed convolution (ConvT) layers, BatchNorm layers and ReLU layers as our image decoder. For 2D and 3D joints, we utilise five stacks of Fully-Connected (Linear) and ReLU layers, and add extra dropout layers between each stacks as our 2D and 3D joints encoders and decoders. For both of our latent hand and modality space discriminators, we use 4 stacks of Fully-Connected (Linear) and ReLU layers with a Sigmoid layer at the end to acquire the normalized probabilities. For our latent modality translator, we also adopt the simple structure with 5 stacks of Fully-Connected (Linear) and ReLU layers. We train our model using ADAM optimizer with the initial learning rate of 10^{-4} and the batch size of 64. The dimensionality of the total latent vector and each sub-latent vector is set to 30 and 15, respectively. Details of selecting the dimensionality will be discussed in Section 4.4.

4.3. Evaluation metrics

There are two popular evaluation metrics available for 3D hand pose estimation: 1) **Mean Joint Error** (*i.e.*, **EPE**) and 2) **Percentage of Correct Keypoints** (*i.e.*, **PCK**).

EPE is the error defined by the average Euclidean distance between estimated and ground truth joint in millimeter (mm). In general, EPE is reported and compared with numbers in tables.

PCK is the joint success rate which is validated by the joints falling in a given threshold range of the Euclidean distance. In general, PCK is reported and compared by the area under the curve (AUC) using figures.

[1]https://github.com/pytorch/pytorch

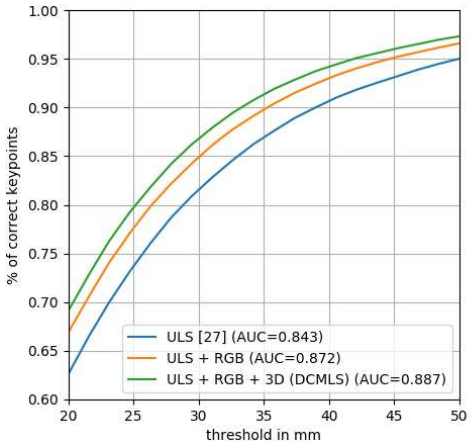

Figure 2. Comparison of 3D PCK performance on the RHD dataset on different baselines models to explore the effectiveness of disentangling modality context in learning a shared hand pose latent space.

γ_z	0.2	0.4	0.5	0.6	0.8
3D EPE	0.861	0.870	**0.887**	0.862	0.860

Table 1. Effectiveness of different latent space capacity for modality specific context and cross-modal hand pose. We compare the 3D PCK on both datasets. A greater γ_z means a larger capacity for cross-modal hand pose latent space.

4.4. Ablation study

Modality context disentanglement. We compare our proposed DCMLS approach with following baselines on the RHD dataset: (1). Unified latent space (ULS) method proposed by [27]; (2). Unified latent space + RGB context disentanglement; (3). Full model: Unified latent space + RGB context disentanglement + 3D context disentanglement. Our results in Figure 3 illustrates and proves that the RGB modality context affects the learning of a modality shared hand pose latent space, hence baseline (2) surpasses baseline (1) from 0.849 to 0.871. While RGB modality context is disentangled, our full model which further disentangles 3D context from 3D modality latent space yields further improvements to achieve the best performance (0.887).

Adversarial modules. We explore different settings for our aforementioned architecture designs: (*i.e.* latent space disentanglement, adversarial learning for cross-modal hand pose space alignment and modality context translation). As summarized in Figure 2, our full model with both Discriminator D_h and D_{3D} in additional to our proposed latent space disentanglement method yields the best result.

Dimensionality of sub-latent vectors. We also discuss

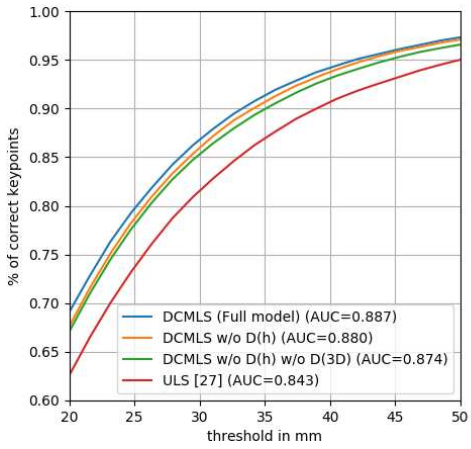

Figure 3. Comparison of 3D PCK performance on the RHD dataset with respect to our proposed architecture designs namely, latent space disentanglement, shared latent space discriminator D_h, context translator discriminator D_{3D}.

the dimensionality of the sub-latent vectors, which is controlled by the capacity ratio of the disentangled representation of hand pose and context. We use γ_z to denote the capacity ratio of the two sub-latent vectors. The greater the ratio, the larger the dimensionality of the hand pose sub-latent vector. Taking the RGB $\rightarrow$ 3D prediction task of hand joints on the RHD dataset as an example, we conduct experiments with the capacity ratio of 0.2, 0.4, 0.5, 0.6 and 0.8, respectively. As shown in Table 1, we observe that when $\gamma_z = 0.5$, the performance of the 3D hand pose estimation is the highest, which is 0.887. As the result, we choose $\gamma_z = 0.5$ for our model in all settings, where the dimensionalities of the sub-latent hand pose and context vectors are set to be the same.

4.5. Performance comparison

We compare our method with several recent studies using both EPE and PCK.

EPE Evaluation. In Table 2, following [43] and [27], we first summarize and compare the results of our DCMLS method with CPose [43], ULS [27], and D-VAE [37] in terms of the average 3D EPE. The second column stands for 3D prediction from 2D key points, which maps the 2D keypoints space into the 3D hand pose configuration on the RHD dataset. The third column predicts 3D keypoints from RGB hand images on the RHD dataset. The fourth column is the 3D keypoints prediction from RGB hand images on the STB dataset. From the results in Table 2 where the figures of existing methods are excerpted from [27], our DCMLS method achieves the best performance on all three tasks using EPE evaluation, which demonstrates the effectiveness of our method on both 2D to 3D and RGB to 3D tasks.

	2D $\rightarrow$ 3D RHD	RGB $\rightarrow$ 3D RHD	RGB $\rightarrow$ 3D STB
CPose [43]	22.43	30.42	8.68
ULS [27]	17.14	19.73	8.56
D-VAE [37]	/	19.95	8.66
Ours	**15.45**	**17.11**	**7.27**

Table 2. Mean EPE comparison with related works.

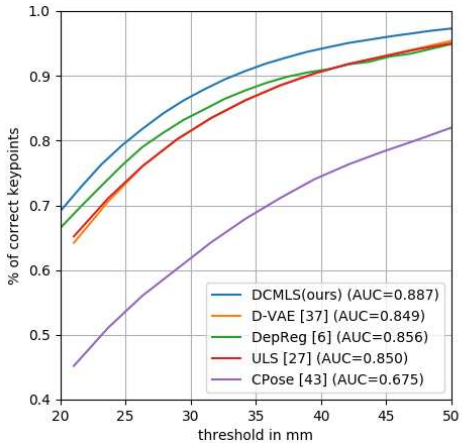

Figure 4. 3D PCK on the RHD dataset

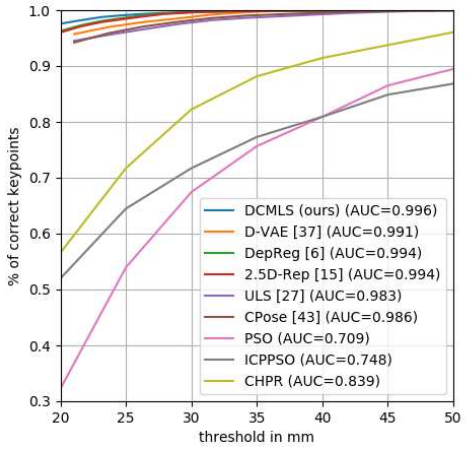

Figure 5. 3D PCK on the STB dataset

PCK Evaluation. In addition, in Figure 4 and Figure 5, we compare our proposed method DCMLS with three more existing works[22, 6, 15] by computing the PCK curve within a fixed thresholds, for the RHD and STB datasets, respectively. As shown in Figure 4, our DCMLS method achieves the best performance among all the methods. In addition, our DCMLS method clearly outperforms our directly related baseline method [27] from 0.983 to 0.996 on the STB dataset and reaches the state-of-the-art performance. Similarly, DCMLS largely surpasses [27] from

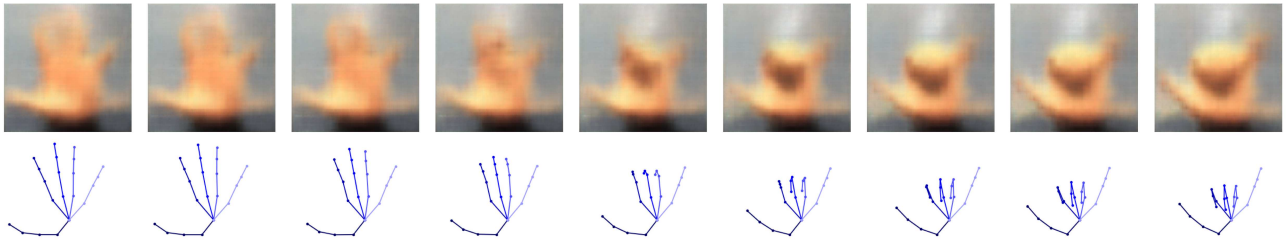

Figure 6. **A latent walk in learnt latent spaces.** The left-most and right-most images are reconstructed from two input images. The interpolated hand pose representation were then used to reconstruct the corresponding RGB images and 3D poses. This shows that our model learns a meaningful and continuous latent space for hand representation.

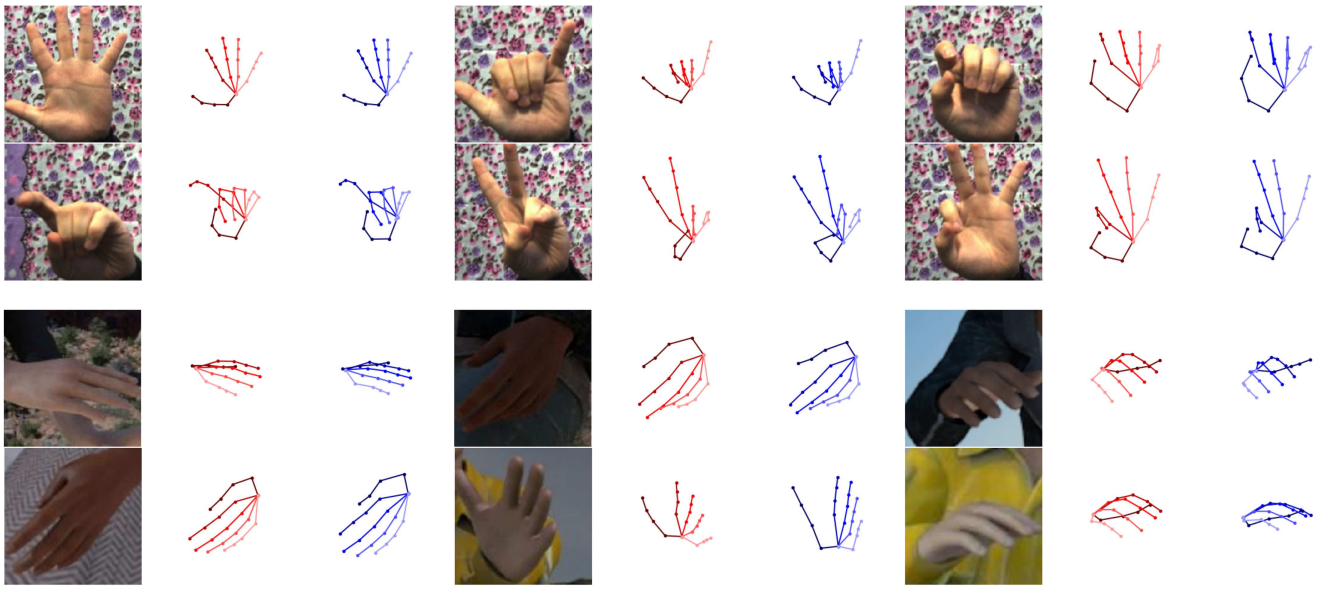

Figure 7. Predicted 3D estimation outputs on STB and RHD. The middle column is our predictions and the right columns are the ground truth.

0.849 to 0.887 on the RHD dataset and reaches the top among all the methods compared. This clearly shows the effectiveness of our disentangled latent spaces for cross modal hand pose estimation.

Qualitative Evaluation. In Figure 6, we show the result of an experiment on a random walk of the hand pose latent space to visualize the synthesized images and corresponding 3D hand poses by using two input image samples (*i.e.*, the left-most and right-most images). We interpolate the in-between representation from the hand pose latent space and keep the context representation fixed. This indicates that meaningful latent representations can be learned through our model. Figure 7 visualizes some predictions from both STB and RHD. Our model predicts precise 3D hand joint locations. Some predictions such as the samples at the 2-th column even show more reasonable or accurate hand pose than the ground truth.

5. Conclusion

In this paper, we present a novel DCMLS model for 3D hand pose estimation from RGB images by distangling a latent space of an input data into modality specific latent space and hand pose specific latent space and formulating the pose estimation task as a hand-pose specific cross-modal learning task. In order to guide the encoders to learn cross-modal features, we add a novel hand pose representation discriminator which learns to distinguish hand features of different modalities while feedback is utilized to improve the encoders so that they encode hand representation without modality context. Meanwhile we make use of the modality context representation to learn a cross-modal translation for pose estination. We conducted comprehensive experiments to demonstrate the improvements of our proposed DCMLS method in learning modality invariant hand representation for improved 3D hand pose estimation.

References

[1] M. Abdi, E. Abbasnejad, C. P. Lim, and S. Nahavandi. 3d hand pose estimation using simulation and partial-supervision with a shared latent space. In *British Machine Vision Conference (BMVC)*, 2018.

[2] V. Athitsos. Estimating 3d hand pose from a cluttered image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2003.

[3] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba. Cross-modal scene networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 40(10):2303–2314, 2018.

[4] S. Baek, K. In Kim, and T.-K. Kim. Augmented skeleton space transfer for depth-based hand pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] A. Boukhayma, R. d. Bem, and P. H. Torr. 3d hand shape and pose from images in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] Y. Cai, L. Ge, J. Cai, and J. Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *European Conference on Computer Vision (ECCV)*, 2018.

[7] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems (NeurIPS)*, pages 2172–2180, 2016.

[8] Y. Chen, Z. Tu, L. Ge, D. Zhang, R. Chen, and J. Yuan. So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[9] M. de La Gorce, D. J. Fleet, and N. Paragios. Model-based 3d hand pose estimation from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 33(9):1793–1805, 2011.

[10] L. Ge, Y. Cai, J. Weng, and J. Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NeurIPS)*, pages 2672–2680, 2014.

[12] A. Harsh Jha, S. Anand, M. Singh, and V. Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–820, 2018.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] J. Hoffman, S. Gupta, J. Leong, S. Guadarrama, and T. Darrell. Cross-modal adaptation for rgb-d detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[15] U. Iqbal, P. Molchanov, T. Breuel, J. Gall, and J. Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *European Conference on Computer Vision (ECCV)*, 2018.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *arXiv e-prints*, page arXiv:1312.6114, Dec 2013.

[18] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision (ECCV)*, September 2018.

[19] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems (NeurIPS)*, pages 700–708, 2017.

[20] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv e-prints*, page arXiv:1411.1784, Nov 2014.

[21] G. Moon, J. Yong Chang, and K. Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[22] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[23] M. Oberweger, G. Riegler, P. Wohlhart, and V. Lepetit. Efficiently creating 3d training data for fine hand pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] M. Oberweger, P. Wohlhart, and V. Lepetit. Generalized feedback loop for joint hand-object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, pages 1–1, 2019.

[25] P. Panteleris, I. Oikonomidis, and A. Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018.

[26] G. Poier, D. Schinagl, and H. Bischof. Learning pose specific representations by predicting different views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[28] B. Stenger, P. R. S. Mendonca, and R. Cipolla. Model-based 3d tracking of an articulated hand. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Dec 2001.

[29] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[30] D. Tang, J. Taylor, P. Kohli, C. Keskin, T.-K. Kim, and J. Shotton. Opening the black box: Hierarchical sampling

optimization for estimating human hand pose. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[31] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):169:1–169:10, Sept. 2014.

[32] C. Wan, T. Probst, L. V. Gool, and A. Yao. Self-supervised 3d hand pose estimation through training by fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10853–10862, 2019.

[33] C. Wan, T. Probst, L. Van Gool, and A. Yao. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[34] C. Wan, T. Probst, L. Van Gool, and A. Yao. Dense 3d regression for hand pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[35] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[36] L. Yang, S. Li, D. Lee, and A. Yao. Aligning latent spaces for 3d hand pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[37] L. Yang and A. Yao. Disentangling latent hands for image synthesis and pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[38] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Yong Chang, K. Mu Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge, J. Yuan, X. Chen, G. Wang, F. Yang, K. Akiyama, Y. Wu, Q. Wan, M. Madadi, S. Escalera, S. Li, D. Lee, I. Oikonomidis, A. Argyros, and T.-K. Kim. Depth-based 3d hand pose estimation: From current achievements to future goals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[39] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv e-prints*, page arXiv:1610.07214, 2016.

[40] W. Zhang, W. Ouyang, W. Li, and D. Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[41] W. Zhang, D. Xu, W. Ouyang, and W. Li. Self-paced collaborative and adversarial network for unsupervised domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2019.

[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[43] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.