Pattern Analysis and Applications (2022) 25:157–167
https://doi.org/10.1007/s10044-021-01048-x

THEORETICAL ADVANCES

3D hand pose estimation from a single RGB image through semantic decomposition of VAE latent space

Xinru Guo[1] · Song Xu[1] · Xiangbo Lin[1] · Yi Sun[1] · Xiaohong Ma[1]

Received: 25 May 2021 / Accepted: 13 December 2021 / Published online: 24 January 2022

Abstract
Based on the disentanglement representation learning theory and the cross-modal variational autoencoder (VAE) model, we derive a "*Single Input Multiple Output*" (SIMO) disentangled model cmSIMO − β VAE. With the guidance of this derived model, we design a new VAE network, named da-VAE, for the challenging task of 3D hand pose estimation from a single RGB image. The designed da-VAE network has a multi-head encoder with the attention modules. Cooperating with the specific supervisions, the latent space is decomposed into subspaces with explicit semantics, which are relevant to the generative factors of hand pose, shape, appearance and others. The performance of the proposed da-VAE network is evaluated on RHD and STB dataset. The experimental results show competitive accuracies with the state-of-the-art methods.

Keywords Hand pose estimation · Monocular RGB image · Disentanglement representation learning · Variational autoencoder

1 Introduction

Vision-based 3D hand pose estimation has fundamental important role in many applications including augmented reality and robotics. Being the most easily obtained visual data with lower cost, single RGB image has been paid much more attentions in estimating 3D hand pose task. Because of the great development of deep learning, significant progress has been made in recent years [1–8]. However, compared to the depth map, estimating the 3D coordinates of each hand joints from a single RGB image is more difficult due to the lack of the depth information, cluttered background, occlusions, and complex illuminations. Faced to this less well solved issue, further researches from different aspects should be encouraged.

RGB images can be considered as the result of a generative process involving many factors of independent/dependent variation. Take a hand image as an example, the hand's pose, shape (here means the geometric attributes, such as the bone length), appearance (color/texture) are generated by independent variables. If the generative factors can be independently encoded from high-dimensional data in a generative model, the downstream tasks would have better solutions [9, 10]. The disentangled representation learning is one of the factor isolation technologies, and the current promising unsupervised disentanglement methods are largely based on the variational autoencoders (VAEs) [11–15]. Recently, a particular appealing work [16] firstly introduced a disentanglement VAE framework into the single RGB image-based hand pose estimation task and obtained promising results. Our research is inspired by this work and provides a new disentanglement VAE structure for hand pose estimation.

By detailed theoretical and experimental analysis, Locatello et al. [17] highlights the role of the inductive bias and the supervisions in disentanglement learning. In the meanwhile, Vahdat et al. [18] claims that VAEs can benefit from designing special network architectures as they have fundamentally different requirements, and powerful encoders can yield better disentanglement. Considering the findings of most research papers relevant to unsupervised disentanglement learning, our work focuses on designing a new 3D hand pose estimation network with an attention module

Xinru Guo and Song Xu contributed equally to this study.

This work was supported by the National Natural Science Foundation of China [grant numbers 61873046, U1708263].

✉ Xiangbo Lin
linxbo@dlut.edu.cn

[1] School of Information and Communication Engineering, Dalian University of Technology, Dalian, China

embedded encoder to get highly accurate estimation. Our main contributions can be summarized as follows.

(1) Based on the unsupervised disentanglement representation learning within the variational autoencoder framework, we derive a "*Single Input Multiple Output*" (SIMO) disentangled model $cmSIMO-\beta\ VAE$. This model provides the theoretical support to separate the hand pose factor, shape factor and other factors for generating a hand image.
(2) According to the $cmSIMO-\beta\ VAE$ model, we design a new hand pose estimation network named da-VAE. It has three attention module embedded heads in the encoder to extract the pose factor effectively when cooperating with the specific supervisions and the weighted KL-divergence.
(3) We have explored the contributions of the network architecture and the influence of the parameter β by experiments and have performed comparisons with many of the state-of-the-art methods, 9 methods on RHD dataset [1] and 15 methods on STB dataset [19].

2 Related works

2.1 Variational autoencoders based disentanglement

As is stated [17], state-of-the-art approaches for unsupervised disentanglement learning are largely based on variational autoencoders (VAEs) [11]. Several variations of VAEs mainly realize disentanglement by enforcing the independence in the latent variables, which originally derives from the promising work of β-VAE [20, 21]. Its studies reveal that a hyper-parameter $\beta > 1$ weighting on the KL-divergence term in the loss function of VAEs will encourage the model to learn independent latent factors. After decomposition of the KL-divergence, it finds that the total correlation (TC) of the latent variables contributes most to the disentanglement purpose. After that, Factor-VAE [22] minimizes this TC term through adopting a discriminator in the latent space with density-ratio trick. β-TCVAE [15] employs a mini-batch weighted sampling for the TC estimation. Kumar et al. [23] uses the moment matching to penalize the divergence between the aggregated posterior and the prior.

Giving the latent variables reasonable semantic definitions is new suggestion for unsupervised disentanglement in present works. JointVAE [24] associates the latent space with the continuous and the categorical factors of variation. BF-VAE [14] emphasizes the importance of partitioning and treating in a different manner the latent dimensions corresponding to relevant factors and nuisances for a good disentanglement.

There have been attempts in literature toward building hybrid models of VAE and GAN on disentanglement. For example, ID-GAN [25] learns disentangled representation using VAE-based models and distills the learned representation with an additional nuisance variable to the separate GAN-based generator for high-fidelity synthesis.

Although extensive efforts have achieved a lot in unsupervised disentanglement learning within VAE framework, different views hold that unsupervised methods do not allow to consistently learn disentangled representations without inductive bias or supervision [17]. Well designed network structure can yield better disentanglement [18]. Earlier work [26] generalizes the standard VAEs by employing a general partially specified graphical model structure in the encoder and decoder to build semi-supervised model for better disentangled representations. Ruiz et al. [27] proposes a reference-based VAEs to exploit the weak-supervision provided by the reference set containing images where the factors of interest are constant. Recently, Chen et al. [28] introduces weak supervision by providing similarities between instances based on a factor to be disentangled. Locatello et al. [29] studies over 52000 models and claims that better disentangled representations can be obtained if the labels are incorporated into the learning process using a simple supervised regularizer.

2.2 Hand pose estimation using VAE framework

Crossing-Net [30] is an early representative approach that use the VAE to model a prior distribution on hand pose configurations. It learns a shared latent space between the depth map and the pose, and integrates the generative adversarial network (GAN) for modeling the distributions of depth images. Gao et al. [31] adopts two VAEs for hand segmentation and 3D canonical and relative hand joint coordinates regression in the scene of a hand grasping objects. Spurr et al proposes the Cross-modal VAEs [32] that leads to a coherent latent space across multiple modalities for hand pose estimation from a single RGB image. To learn a joint latent representation, Yang et al. [33] adopts latent space alignment from individual modalities to leverage other modalities as weak labels for a better RGB-based hand pose estimation. In addition, Kulon et al. [34] proposes the AE model consisting of an image encoder followed by a mesh convolutional decoder to directly reconstruct the hand mesh and estimate the 3D hand pose.

Different from the above mentioned methods, d-VAE [16] firstly uses the independent factors of variations to learn disentangled representations for 3D hand pose estimation. The used interpretable factors of variations are 3D pose, canonical pose and camera viewpoint.

For derivation, it adopts a separated disentangling step and an embedding step to get a proper latent space decomposition.

3 Da-VAE network

3.1 *cmSIMO* − *β VAE* model

Up to present, there is no widely accepted definition of latent space disentanglement [17], but all the definitions agree that a disentangled latent space representation should separate the distinct, informative factors of variations in the data [9]. We can take a single RGB image as an example to give a more straightforward explanation. A single RGB image is a complex products relevant to many factors. The latent vector z is the representation of these factors. If each component z_j in z is statistically correlated with only a single factor, we say that the latent vector z is disentangled. That is, varying one z_j while fixing other components will just change the image contents relevant to the *j–th* factor.

In vision perception applications, for a specific task, only limited image contents are needed to be analyzed. Those irrelevant contents should be isolated or discarded. As the expectation of the disentanglement, the relevant contents could be embedded into the latent space, forming the latent vector z with the independent components z_j. These independent components should have explicit semantics and should be related to the generation factors. The irrelevant contents have no or negative impacts on the task. They can be preserved or discarded, depending on the need of supervisions.

According to the Cross-modal VAEs [32] which proves that the VAE framework has the ability to project any modality into a lower dimensional latent space z and to generate posterior estimation in any modality, we derive the ELBO $L_{\text{cmSIMO-VAE}}(\Theta, \phi; x, y, z)$ of the proposed "*Single Input Multiple Output*" (SIMO) disentangled model cmSIMO -VAE. Its ELBO is given as follows:

$$\begin{aligned} L_{\text{cmSIMO -VAE}}(\Theta, \phi; x, y, z) &= \sum_{i=1}^{n} L(\theta_i, \phi; x, y_i, z) \\ &= \sum_{i=1}^{n} \alpha_i \mathbb{E}_{q(z|x)} \log p(y_i|z) - KL(q(z|x)||p(z)) \end{aligned} \tag{1}$$

where x is the input data, $y = [y_i,\ i = 1,\ \cdots,\ n]$ is the observation set,ϕ and Θ are the parameter set of the encoders and decoders, α_i are adjustable hyper-parameters to balance the reconstruction quality and the similarity between the empirical posterior probability and the prior latent space distribution $p(z) = \mathcal{N}(0, I)$ is a Gaussian prior on the latent space.

According to the presented notion of the disentanglement, we suggest the principle of the latent space decomposition being the explicit semantic representation and the independence of each latent component. We decompose the latent variables z into n independent components $z = \left[z_{y_1}, z_{y_2}, \ldots z_{y_n}\right]$. Each z_{y_i} is associated with a specific observation y_i and is interpretable. Borrowing the β-trick from β-VAE [20], we can further derive the disentangled model cmSIMO − βVAE . Its ELBO is given as follows:

$$\begin{aligned} L_{\text{cmSIMO}-\beta\,\text{VAE}}(\Theta, \phi; x, y, z) &= \sum_{i=1}^{n} L_{\text{cm}-\beta_i\,\text{VAE}}(\theta_i, \phi; x, y_i, z_i) \\ &= \sum_{i=1}^{n} \alpha_i \mathbb{E}_{q(z_i|x)} \log p(y_i|z_i) \\ &\quad - \beta_i KL(q(z_i|x)||p(z_i)) \end{aligned} \tag{2}$$

where the adjustable parameter $\beta = \left[\beta_i, i = 1, \ldots n\right]$.

Our goal is to estimate the 3D hand pose from a RGB image, where the articulated hand pose, hand geometric shape, hand appearance (color/texture) are independent attributes of a hand RGB image. So we set $n = 3$ of the *cmSIMO* − *βVAE* . The three modalities y_1, y_2, y_3, respectively, represent the 3D hand pose, 3D point cloud of a hand and the hand RGB image. The overall loss for training is given as follows:

$$\begin{aligned} L_{\text{overall}} &= \alpha_1 L_{\text{pose}} + \beta_1 KL_{\text{pose}} + \alpha_2 L_{\text{cloud}} \\ &\quad + \beta_2 KL_{\text{cloud}} + \alpha_3 L_{\text{RGB}} + \beta_3 KL_{\text{RGB}} \end{aligned} \tag{3}$$

where the L2 loss between the prediction and the ground truth is used in L_{pose} and L_{RGB}. To reconstruct the point clouds, we take the Chamfer distance as the similarity metric L_{cloud}. Additionally, the three sub-latent variable distributions are, respectively, aligned with the Gaussian prior distribution by minimizing their KL divergence loss.

3.2 Network architecture

Based on the theoretical analysis presented in Section 3.1, we build a new attention embedded disentanglement VAEs architecture for the task of estimating the 3D hand pose from a single RGB image, as is shown in Fig. 1.

The articulated hand pose, hand geometric shape, hand appearance (color/texture) are independent attributes of a hand. Together with other factors such as the background, illuminations, etc., they can form a natural RGB hand image. The main goal of our work is to estimate the 3D coordinates of the hand joints. If the relevant generation factor could be separated from other factors, more accurate results would be obtained. To meet this expectation, our proposed da-VAE network has one encoder and three decoders. The encoder is split

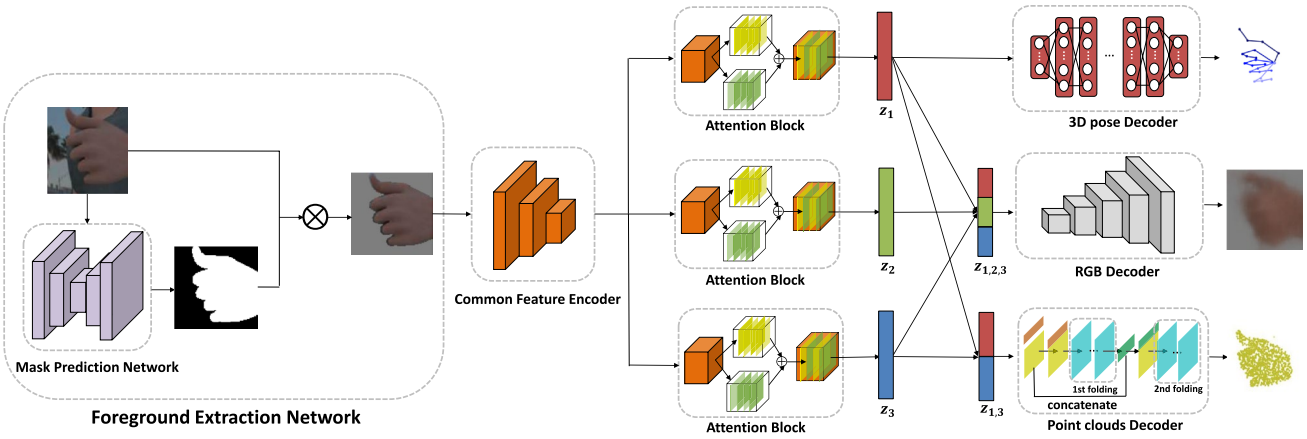

Fig. 1 Overview of our proposed attention embedded disentanglement VAEs (da-VAE) model

into two part. Specifically, we use the residual structure similar to ResNet-34 [35] in the former part to extract the shared basic features and SKNet [36] in each branch of the latter part to extract the output-related specific features. The three decoders are used to reconstruct the different components of a hand image. These components might have different modalities.

The encoder generates the latent variable z and disentangles it into three sub-latent variables. We set the dimensionality of the total latent variable z to 48 and the disentangled three sub-latent variables z_1,z_2 and z_3 to 24, 16 and 8, respectively. We predict the 3D joint locations by decoding the sub-latent variable z_1 . The decoder for the 3D joint locations consists of stacks of fully connected and ReLU layers. We construct the hand shape related latent variable $z_{1,3}$ by concatenating the sub-latent variable z_1 and z_3 , then decode $z_{1,3}$ to generate the articulated hand shape. The hand shape is represented by the point clouds. We adopt the FoldingNet architecture [37] as the point clouds decoder. For hand RGB image reconstruction, the decoder contains several transposed convolution, BatchNorm and ReLU layers, and decodes the total latent variable z.

In addition, in order to alleviate the influence of the cluttering background, we use ResNet-34 [35] to learn a hand mask extraction model utilizing the provided hand segmentation mask. The hand mask model is pre-trained and its weights are frozen during training the multi-head attention embedded disentanglement VAEs model.

3.3 Attention mechanism

Attention can be interpreted as a means of biasing the allocation of available computational resources toward the most informative components of a signal [38]. The basic idea of the attention mechanism in computer vision is to make the system have the attention ability to focus on key information, while to ignore irrelevant information. It helps the model to search for more powerful representations. Our da-VAE network aims to decompose the encoded hidden space into three subspace endowed with semantic meaning of 3D hand pose, hand shape and hand color/texture respectively, which correspond to the generation factors of a hand color image. Adopting the attention mechanism to construct the encoder should be useful to selectively enhance those informative features while to suppress those interference for the specific generation factors. To achieve this goal, we use the SKNet [36] network structure as the latter part of each head in the encoder. SKNet proposes a dynamic selection mechanism in CNNs that allows each neuron to adaptively adjust its receptive field size based on multiple scales of input information. Specifically, SKNet proposes a "Selective Kernel" (SK) convolution, which mainly includes three steps: SPLIT, FUSE and SELECT. The SPLIT step uses convolution kernels of different size in multiple branches to form the features with varied receptive fields. The FUSE step aims to obtain a global representation for selection weights by aggregating information from multiple branches. According to the selection weights on various kernels, the SELECT step uses the soft attention across channels to obtain final feature map by aggregating features at different scales. SKNet is an imitation of the behavior of the human visual cortical neurons and can provide selective attention caused by the specific stimulus. In our work, the stimulus comes from the hand pose, shape and hand image to be reconstructed.

4 Experiments and results

4.1 Datasets and evaluation metrics

The evaluation experiments use two public datasets for 3D hand pose estimation: the Rendered Hand Pose Dataset

(RHD) [1] and the Stereo Hand Pose Tracking Benchmark (STB) [19].

RHD Dataset RHD is a synthesized dataset of rendered hand images built upon 20 different subjects performing 39 actions with diverse hand sizes. It contains 41258 training images and 2728 test images with a resolution of 320×320 pixels. For each RGB image, 2D and 3D joint locations are provided, so are the corresponding depth map and segmentation mask. This dataset is extremely challenging due to the large variations in viewpoints, as well as different illuminations and various hand poses.

STB Dataset STB is an earlier released, widely used and relatively simple dataset for 3D hand pose estimation from RGB images. It has 18k stereo pairs captured with a multi-view binocular camera setup, where 15k images for training and 3k images for testing. Since the annotations are acquired manually, only a single person's left hand in front of 6 real-world indoor backgrounds with variant illuminations is recorded, and most regions of the hands are visible. The resolution of each image is 640×480. The annotations include 3D joint positions of palm and fingers, and the camera intrinsic configurations.

Compared with the STB dataset, the RHD dataset has more data, richer gestures, more cluttered backgrounds, more viewpoints and more complex illuminations. So we use the RHD dataset to perform the following studies about the parameter selection and the latent space decomposition analysis.

Evaluation Metrics There are three common evaluation metrics available for evaluating the performance of our proposed method and comparing with other works on 3D hand pose estimation:

(1) *Mean end-point-error (EPE)* which measures the average Euclidean distance in millimeters (mm) between the predicted 3D joints and the ground truth joints.
(2) *Percentage of Correct Keypoints (PCK)* which is defined by the percentage of predicted keypoints that fall within a given threshold range of the Euclidean distance compared with the corresponding ground truth.
(3) *Area Under Curve (AUC)* which is the measure of the area enclosed by the PCK curve.

4.2 Implementations

Data Pre-processing We crop out the hand area for each RGB image and resize it to 256×256. Using the camera intrinsic parameters provided by the dataset, the corresponding hand area in the depth image is converted to the point clouds for supervision of the output shape modality. After cropping the hand from the RGB image, we perform view normalization on the 3D joints and the corresponding point clouds according to the method in [39] to rotate the 3D hand so that the center of the 3D hand can be aligned with the z-axis of the camera.

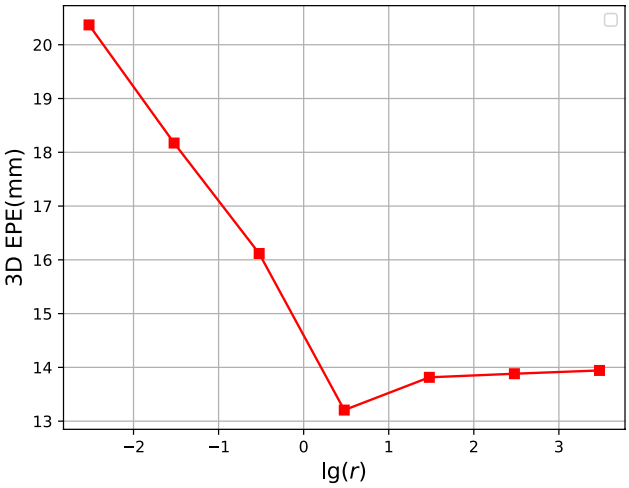

Fig. 2 the curve of 3D end-point-error EPE *vs*. lg(r)

Data augmentation We randomly augment the hand scale between [0.8, 1] and rotation them between $[\pi, -\pi]$ around camera view axis. The corresponding 3D pose labels and 3D point clouds are also rotated accordingly.

Training Our model is implemented with Pytorch framework. We choose ADAM as our optimizer with the initial learning rate of 10^{-3} and the mini-batch of 32. We apply learning rate decay strategy to lower the learning rate from 10^{-4} to 10^{-7}.

4.3 Parameter selection

To get better weighting parameters of the loss function, detailed experiments are carried out by using RHD dataset. First, let $\alpha_1, \alpha_2, \beta_1 = \beta_2 = \beta_3 = \beta$, be invariant positive numbers, $r = \beta/\alpha_3$. Changing the ratio r from 0.003 to 3000, we can get the curve of the mean 3D end-point-error EPE on the RHD test dataset from the trained da-VAE model, shown in Fig. 2. The experimental results indicate that choosing the values of r in the range of 3∼30 will produce smaller 3D end-point-error.

Then, leave the current values of α_3 and β unchanged, and optimize the parameters α_1 and α_2. The influence of the parameters α_i on the 3D end-point-error EPE is shown in Fig. 3, where the horizontal axis is set as the logarithm of α_i/α_3 . It can be seen that when $\alpha_1/\alpha_3 \geq 1000$ and $\alpha_2/\alpha_3 \geq 100$, the 3D end-point-error EPE is minimized.

4.4 Attention mechanism study

We explore the contributions of the attention mechanism through comparing two models on RHD dataset: (1)the baseline model **na-VAE**; (2)the proposed model **da-VAE**.

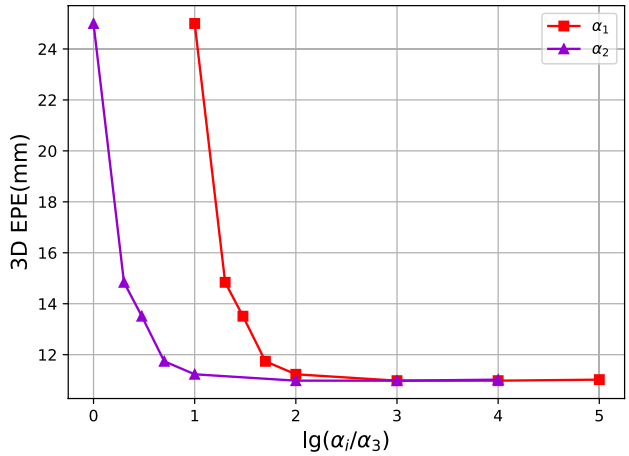

Fig. 3 The influence of the parameters α_i on the 3D end-point-error EPE

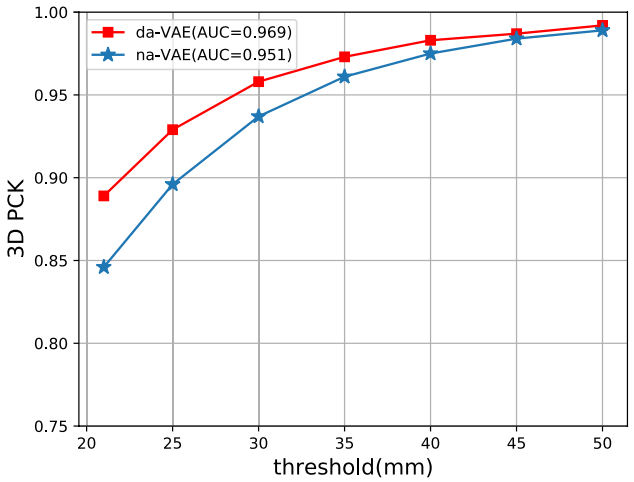

Fig. 4 The comparisons of PCK metric of the network architecture with and without attention mechanism on RHD dataset. The AUC values are shown in parentheses

Table 1 The comparisons of the hand pose estimation results on network architecture with and without attention mechanism on RHD dataset. The symbol '↓' in EPE metric means lower is better, while the symbol '↑' in AUC metric means higher is better

	EPE(mm)↓	AUC ↑
na-VAE	12.56	0.951
da-VAE	10.96	0.969

Similar to the da-VAE model, the encoder of the na-VAE model is also split into two parts. The structure of the encoder's former part is the same as the da-VAE model, but in the latter part, the three attention blocks in the da-VAE model are replaced with common ResNet-34 [35] structures without using any attention mechanism. The decoder is the same as the proposed da-VAE model. The parameters of the two models are equivalent. The main goal of our task is to estimate the 3D coordinates of the hand joints. We perform performance comparison using EPE, 3D PCK and AUC metrics on RHD dataset in Table 1 and Fig. 4. It can be seen that the results of the proposed da-VAE network architecture outperform na-VAE. Compared with na-VAE model, the obtained EPE metric of the da-VAE model has a 1.60mm reduction, and the AUC metric and PCK curves have obvious improvements. The experimental results indicate that the da-VAE can extract more effective latent variables for the task of hand pose estimation due to the adopted attention mechanism in the encoder. The attention mechanism is conducive to decompose the pose factor by aggregating information from multiple kernels to realize the adaptive receptive field sizes of neurons. In addition, we also compare two models using quantitative metrics for the latent embedding disentanglement including the ELBO and the KL of three latent subspace metric in Table 2. From Table 2, we can see both the na-VAE and the da-VAE have small KL divergences on 3D pose, 3D point cloud and the hand RGB branches. The smaller the KL divergence, the better the matching between the real distribution and the approximate distribution. In other words, the na-VAE model and the proposed da-VAE model make a good latent space disentanglement. Compared with na-VAE model, the da-VAE model has better disentanglement ability, because its KL divergences are smaller, and the ELBO of da-VAE is − 12.628, larger than −17.618 of the na-VAE model.

Table 2 The comparisons of ELBO and KL metric of the network architecture with and without attention mechanism on RHD dataset

	ELBO	KL_{pose}	KL_{cloud}	KL_{RGB}
na-VAE	− 17.618	4.533	2.884	2.844
da-VAE	− 12.628	3.264	1.959	1.990

According to the above experimental results, it can be concluded that the attention mechanism can lead to better disentanglement of the latent space. The EPE, 3D PCK and AUC metrics focus on evaluating the performance of the downstream 3D hand pose estimation, which indirectly reflects the disentanglement quality of the latent space. The experimental results in Table 1 and Fig. 4 have proved that the da-VAE model with the attention mechanism is superior to the na-VAE model without the attention mechanism. The ELBO and KL divergence metrics focus on directly evaluating the disentanglement quality of the latent space. The experimental results in Table 2 also prove that the proposed da-VAE model is better than the na-VAE model. The attention mechanism indeed helps the learned latent subspace represent more powerful corresponding semantic

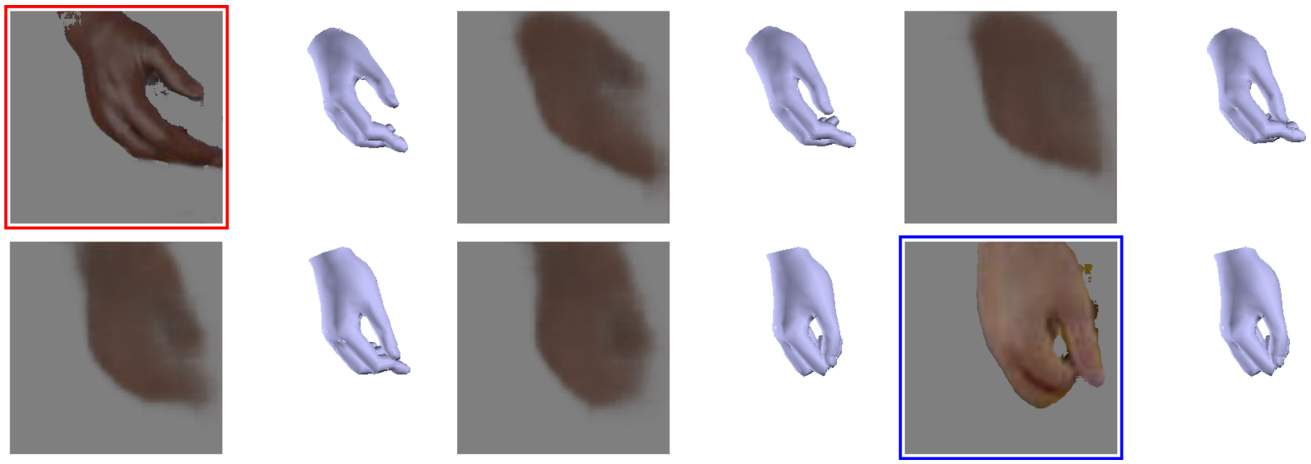

Fig. 5 The result of changing z_{y_1} with invariant z_{y_2},z_{y_3}. In each group, the left image is the reconstructed hand image, and the right is regressed hand pose. The images in the red and blue box, respectively, are the start and end reference image

information through enhancing useful features and suppressing irrelevant features for specific tasks.

4.5 Visual evaluation

According to our proposed da-VAE network, the latent space is decomposed into three subspace z_{y_1}, z_{y_2}, z_{y_3}, endowed with semantic meaning of 3D hand pose, hand color/texture factors and hand shape, respectively. To observe whether each subspace really conforms to the given semantics, we select two test images randomly as the start and end reference image, corresponding to the images in the red and blue box, respectively, in Figs. 5 and 6. First, keeping the subspaces $z^0_{y_2}$, $z^1_{y_2}$ and $z^0_{y_3}$, $z^1_{y_3}$ of the start and end reference image unchanged, z_{y_1} is modified by several linear interpolations between $z^0_{y_1}$ and $z^1_{y_1}$. As expected, the hand pose changes from the start to end reference image gradually without significant impacts on the hand shape and the hand color/texture. A typical result is shown in Fig. 5, where the hand poses are displayed with rendered mesh model for better visualization. The mesh model employs the parametric hand model, MANO [40]. We use the BiHand [41] model to map the predicted 3D joint coordinates into the MANO's parameters. These parameters are then used to recover the hand mesh model.

Second, keeping the subspaces $z^0_{y_1}$, $z^1_{y_1}$ and $z^0_{y_3}$, $z^1_{y_3}$ of the start and end reference image unchanged, z_{y_2} is modified by several linear interpolations between $z^0_{y_2}$ and $z^1_{y_2}$. It can be seen from Fig. 6 that the hand color changes from the

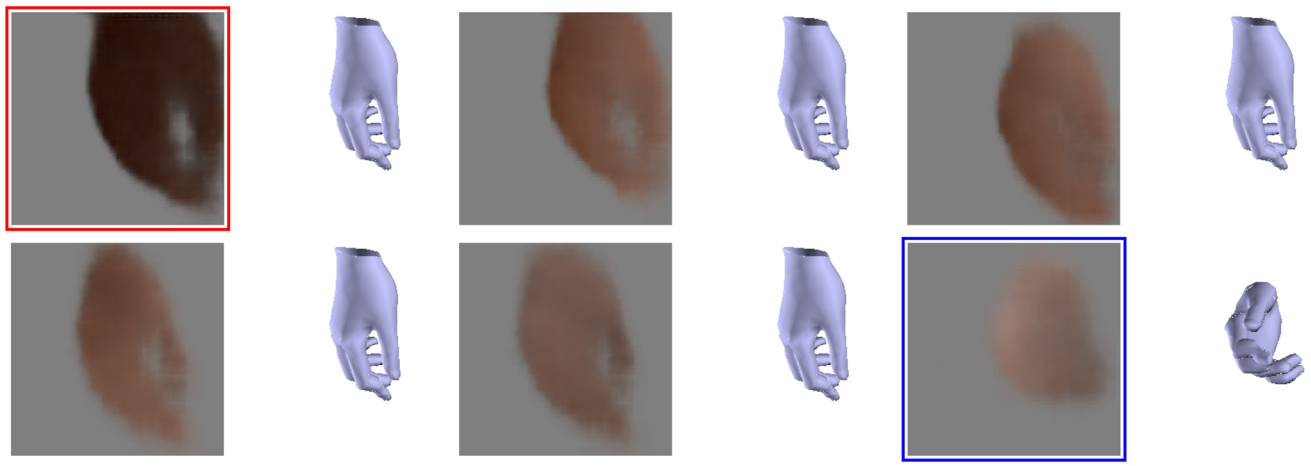

Fig. 6 The result of changing z_{y_2} with invariant z_{y_1},z_{y_3}. In each group, the left image is the reconstructed hand image, and the right is regressed hand pose. The images in the red and blue box, respectively, are the start and end reference image

start to end reference image gradually without significant impacts on the hand shape and the hand pose.

Similar analysis is carried out on the subspace z_{y_3}. Due to the lack of hand shape diversity in the dataset, there is little obvious difference in reconstruction results. But as expected, the hand pose and the hand color/texture are no obvious changes.

These experimental results indicate that by decomposing the hidden space at the semantic level of pose, shape and color/texture, the proposed da-VAE network can achieve the goal of separating the related generation factors of a hand color image.

4.6 Compared with the state-of-the-art methods

We compare our da-VAE model with some other state-of-the-art methods [1–3, 5–7, 16, 32–34] on RHD dataset. We divide these methods into three groups according to the characteristics of the model framework. The first group consists of those models using the VAE framework [16, 32, 33]. The second group focuses on the hand mesh recovery task, accompanied by the hand pose estimation task [5–7, 34]. The third group includes other typical hand pose estimation model using other neural network frameworks [1–3].

Specifically, in the first group, Spurr et al. [32] adopts a multiple-encoders/ multiple-decoders VAE architecture with a shared latent space, since it has been proved that different modalities have a unified latent space. After analyzing that the joint posterior is proportional to the product of individual posteriors, Yang et al. [33] adopts multiple VAEs architecture to get multiple unimodal latent representations, then estimates the joint latent representation by latent space alignment via the "*Gaussian Experts*". The designed VAEs framework in [16] consists of three encoders and three decoders dealing with the data of different modalities and performing the latent space disentanglement. It uses a special two-stage training strategy to decouple the learning of disentangling factors and the embedding of image content. Our proposed **da−VAE** model has one encoder and three decoders, where the encoder is split into two parts. The former part uses the conventional residual structure to extract the shared features among different targets, while the latter part is split into three heads. Each head is equipped with a special soft attention module to help disentangling the latent variable.

In the second group, the network framework in [5] integrates the stacked hourglass network, the residual network and the Graph CNN. Zhang et al. [6] first utilizes the Stacked Hourglass network for 2D hand pose estimation, then uses a simple fully convolutional and multiple fully connected layers to iteratively regress the camera parameters and the mesh parameters. Baek et al. [7] uses the autoencoder (AE) network to estimate the 2D skeletal joints, then uses a single fully connected layer to estimate the 63-dimensional mesh parameter and performs the iterative mesh refinement via back-projection. Kulon et al. [34] adopts an AE structure for monocular 3D hand pose estimation and mesh recovery, where the ResNet-50 is used as the encoder and the spiral patch operator is embedded into the mesh convolutional decoder.

In the third group, the framework in [1] adopts the conventional CNN network with multiple convolution/RELU/maxpooling layers and across-layer concatenation. Iqbal et al. [2] uses an AE network architecture with skip connections for 2.5D heatmap regression, then converts the 2.5D heatmap to the corresponding 3D coordinates by a soft-argmax operation. Cai et al. [3] uses the fully convolutional AE network to get the 2D joints heatmaps, then infers the depth of each joint with two convolutional layers and three fully connected layers, and finally uses a deconvolution network with two convolutional layers and three fully connected layers to render the corresponding depth map.

The results of different methods on RHD dataset are shown in Table 3. From the results of Group 1, it can be seen that the adopted "*Gaussian Experts*" strategy in the Latent Align method [33] provides better latent space alignment between different modalities, which brings an error reduction of more than 6mm in comparison with CM-VAE [32] and d-VAE [16] methods. Our proposed **da−VAE** method yields about 8.99mm error reduction, the best result in this group. The AUC metric is also improved about 14%. All the methods in Group 1 belong to VAE framework, but with different architecture considerations. Our da-VAE architecture is superior to the other three methods.

Since the methods in the GROUP 2 mainly study the hand mesh recovery problem, most of them only report the AUC metric on hand pose estimation problem. The exception is Kulon et al. [34]. It presents the hand pose estimation error, less than 0.04mm than our result. However, our AUC metric is slightly better. Therefore, the performance of the two methods is on par.

The best results obtained using the methods in the GROUP 3 are 13.41mm in EPE metric and 0.94 in AUC metric, which have 2.45mm more error and 3% lower AUC than our results.

In summary, from Table 3, it can be proved that when taking as input the single RGB image, our architectural design is better than most of the state-of-the-art models on dealing with the challenging RHD dataset.

We also compare our da-VAE model on STB dataset with fifteen state-of-the-art methods [1–8, 16, 31–33, 42–44]. Since STB dataset is released earlier, the backgrounds are relatively simple and the hand poses are not complex. We can see from Table 4 that most methods have an AUC metric larger than 99%, which is already saturated. It should be noted that the reported AUC metric computes the area of the PCK curve in

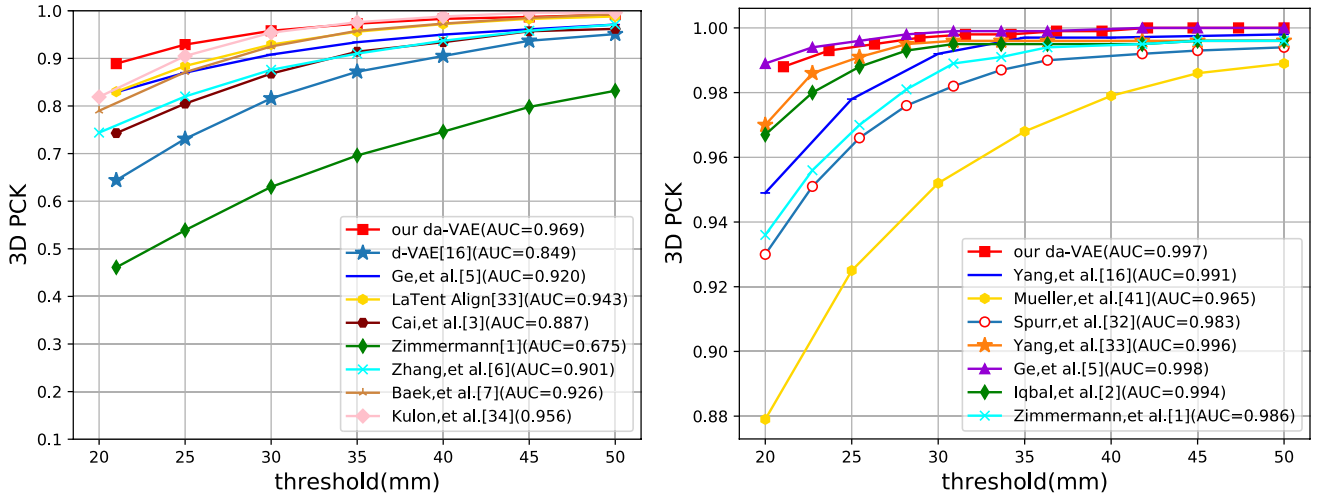

Fig. 7 The comparisons of PCK metric of different models on RHD dataset (left) and STB dataset (right). The AUC values are shown in parentheses

Table 3 The comparisons of the hand pose estimation results on RHD dataset. The symbol '↓' in EPE metric means lower is better, while the symbol '↑' in AUC metric means higher is better. The symbol '*[.]' is used to note that the value is got from the literature annotated after the asterisk. The symbol '–' is used to note that the result is not reported. We highlight our results in bold

GROUP	METHOD	EPE (mm) ↓	AUC ↑
1	CM-VAE [32]	19.73	0.849
	d-VAE [16]	19.95	0.849
	Latent Align [33]	13.14	0.943
	da-VAE	**10.96**	**0.969**
2	Ge et al. [5]	–	0.920
	Zhang et al. [6]	–	0.901
	Baek et al. [7]	–	0.926
	Kulon et al. [34]	10.92	0.956
3	Zimmermann [1]	30.42* [16]	0.675* [42]
	Iqbal [2]	13.41	0.94
	Cai et al. [3]	–	0.887

Table 4 The comparisons of the hand pose estimation results on STB dataset. The symbol '↓' in EPE metric means lower is better, while the symbol '↑' in AUC metric means higher is better. The symbol '*[.]' is used to note that the value is got from the literature annotated after the asterisk. The symbol '–' is used to note that the result is not reported. We highlight our results in bold

METHOD	EPE (mm) ↓	AUC ↑
Ge et al. [5]	6.37	0.998
da-VAE (ours)	**6.61**	**0.997**
Cai et al. [8]	6.95	0.995
Yang et al. [33]	7.05	0.996
Zhao et al. [43]	8.18	0.987
Spurr et al. [32]	8.56	0.983
Yang et al. [16]	8.66	0.991
Zimmermann et al. [1]	8.68* [16]	0.986* [16]
Gao et al. [31]	8.943	0.984
Boukhayma et al. [4]	9.76	0.994* [42]
Zhang et al. [6]	–	0.995
Baek et al. [7]	–	0.995
Iqbal et al. [2]	–	0.994
Cai et al. [3]	–	0.994
Mueller et al. [44]	–	0.965
Zhou et al. [42]	–	0.898

the joint error range between 20mm and 50mm. When the error threshold is set to be less than 20mm, there will be a big drop in the AUC metric. Further efforts should be made to improve the AUC metric when the threshold is below 20mm for certain fine manipulation scenarios such as accurate UI interactions. For the EPE metric, our method outperforms the other 14 methods, 0.24 mm higher than the result of Ge et al. [5].

The PCK curves over different error thresholds from 20mm to 50mm on RHD dataset and STB dataset are shown in Fig. 7. It can be seen that on RHD dataset, as shown in the left in Fig. 7, our da-VAE model outperforms all the compared state-of-the-art models over all the error thresholds. On STB dataset, as shown in the right in Fig. 7, our da-VAE model outperforms almost all the methods over all the error thresholds, and the PCK values are only slightly lower than those of Ge et al. [5] over the error thresholds below 35mm.

Figure 8 shows some visual results of our method on RHD dataset. The selections follow the principle of diversity as far as possible, including representative gestures and scenarios. The results show that our proposed model can reliably handle the challenging poses with various orientations and complicated pose articulation.

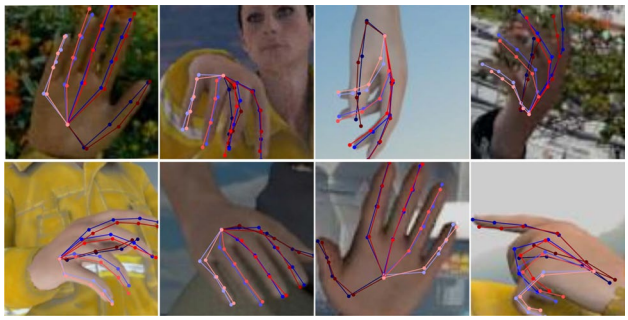

Fig. 8 Examples of the estimated 3D pose on RHD dataset. The red line means the estimation, and the blue line means the ground truth

5 Conclusions

The purpose of this paper is to tackle the challenging task of 3D hand pose estimation from a single RGB image. To extract the unique feature representations closely related to the hand pose from the data, we have studied the unsupervised disentanglement model within the variational autoencoder framework and derived a "*Single Input Multiple Output*" (SIMO) disentangled model $cmSIMO - \beta\ VAE$. With the support of the theory and the knowledge from other literatures, we propose the da-VAE network with three heads in the encoder. The embedded attention module SKNet in each head can imitate the visual attention mechanism and is beneficial to decompose the pose factor when cooperating with the supervisions and the selected weights in the loss function. We perform experiments to explore the effects of the semantic decomposition of the latent space and find that the latent subspaces have the correspondences to the hand pose, shape and color/texture. Comparison experiments on two public datasets verify the superiority of our method.

References

1. Zimmermann C, Brox T (2017) Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, pp. 4903–4911
2. Iqbal U, Molchanov P, Gall TBJ, Kautz J (2018) Hand pose estimation via latent 2.5d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 118–134
3. Cai Y, Ge L, Cai J, Yuan J (2018) Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 666–682
4. Boukhayma A, Bem R de, Torr PHS (2019) 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10843–10852
5. Ge L, Ren Z, Li Y, Xue Z, Wang Y, Cai J, Yuan J (2019) 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10833–10842
6. Zhang X, Li Q, Mo H, Zhang W, Zheng W (2019) End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2354–2364
7. Baek S, Kim KI, Kim TK (2019) Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1067–1076
8. Cai Y, Ge L, Liu J, Cai J, Cham T-J, Yuan J, Thalmann NM (2019) Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2272–2281
9. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
10. Ridgeway K (2016) A survey of inductive biases for factorial representation-learning. *arXiv preprint*arXiv:1612.05299*, 2016*
11. Kingma DP, Welling M (2014) Auto-encoding variational bayes. In *International Conference on Learning Representation (ICLR)*
12. Kulkarni TD, Whitney W, Kohli P, Tenenbaum JB (2015) Deep convolutional inverse graphics network. *Advances in Neural Information Processing Systems (NIPS)*, pp 2539–2547
13. Karaletsos T, Belongie S, Rtsch G (2016) Bayesian representation learning with oracle constraints. In *International Conference on Learning Representations (ICLR)*
14. Kim M, Wang Y, Sahu P, Pavlovic V (2019) Bayes-factor-vae: Hierarchical bayesian deep auto-encoder models for factor disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2979–2987
15. Chen RTQ, Li X, Grosse R, Duvenaud D (2018) Isolating sources of disentanglement in variational autoencoders. *arXiv preprint*arXiv:1802.04942
16. Yang L, Yao A (2019) Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9877–9886
17. Locatello F, Bauer S, Lucic M, Raetsch G, Gelly S, Schölkopf B, Bachem O (2019) Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, pp. 4114–4124
18. Vahdat A, Kautz J (2020) Nvae: a deep hierarchical variational autoencoder. *arXiv preprint*arXiv:2007.03898
19. Zhang J, Jiao J, Chen M, Qu L, Xu X, Yang Q (2016) 3d hand pose tracking and estimation using stereo matching. *arXiv preprint*arXiv:1610.07214
20. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A (2017) β-vae: learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*
21. Burgess CP, Higgins I, Pal A, Matthey L, Watters N, Desjardins G, Lerchner A (2018) Understanding disentangling in β-vae. *arXiv preprint*arXiv:1804.03599
22. Kim H, Mnih A (2018) Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658
23. Kumar A, Sattigeri P, Balakrishnan A (2017) Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations (ICLR)*

24. Dupont E (2018) Learning disentangled joint continuous and discrete representations. *Adv Neural Inf Process Syst (NIPS)*, pp. 710–720
25. Lee W, Kim D, Hong S, Lee H (2020) High-fidelity synthesis with disentangled representation. In *European Conference on Computer Vision (ECCV)*, pp. 157–174
26. Siddharth N, Paige B, van de Meent J-W, Desmaison A, Goodman N, Kohli P, Wood F, Torr P (2017) Learning disentangled representations with semi-supervised deep generative models. Adv Neural Inf Process Syst (NIPS) 30:5925–5935
27. Ruiz A, Martinez O, Binefa X, Verbeek J (2019) Learning disentangled representations with reference-based variational autoencoders. *arXiv preprint*arXiv:1901.08534
28. Chen J, Batmanghelich K (2020) Weakly supervised disentanglement by pairwise similarities. Proce AAAI Conf Artif Intell 34:3495–3502
29. Locatello F, Tschannen M, Bauer S, Rätsch G, Schölkopf B, Bachem O (2019) Disentangling factors of variation using few labels. *arXiv preprint*arXiv:1905.01258
30. Wan C, Probst T, Van Gool L, Yao A (2017) Crossing nets: combining gans and vaes with a shared latent space for hand pose estimation. In *Proc IEEE Conf Computer Vision Pattern Recogn (CVPR)*, pp. 680–689
31. Gao Y, Wang Y, Falco P, Navab N, Tombari F (2019) Variational object-aware 3-d hand pose from a single rgb image. IEEE Robot Autom Letts 4(4):4239–4246
32. Spurr A, Song J, Park S, Hilliges O (2018) Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 89–98
33. Yang L, Li S, Lee D, Yao A (2019) Aligning latent spaces for 3d hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2335–2343
34. Kulon D, Guler RA, Kokkinos I, Bronstein MM, Zafeiriou S (2020) Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4990–5000
35. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016
36. Li X, Wang W, Hu X, Yang J (2019) Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 510–519
37. Yang Y, Feng C, Shen Y, Tian D (2018) Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 206–215
38. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141
39. Li S, Lee D (2019) Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11927–11936
40. Romero J, Tzionas D, Black MJ (2017) Embodied hands: modeling and capturing hands and bodies together. ACM Trans Graph (ToG) 36(6):1–17
41. Yang L, Li J, Xu W, Diao Y, Lu C (2020) Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. *arXiv preprint*arXiv:2008.05079
42. Zhou Y, Habermann M, Xu W, Habibie I, Theobalt C, Xu F (2020) Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5346–5355
43. Zhao L, Peng X, Chen Y, Kapadia M, Metaxas DN (2020) Knowledge as priors: cross-modal knowledge generalization for datasets without superior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6528–6537
44. Mueller F, Bernard F, Sotnychenko O, Mehta D, Sridhar S, Casas D, Theobalt C (2018) Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 49–59