Disentangling Latent Hands for Image Synthesis and Pose Estimation

Linlin Yang
University of Bonn, Germany
yangl@cs.uni-bonn.de

Angela Yao
National University of Singapore, Singapore
ayao@comp.nus.edu.sg

Abstract

Hand image synthesis and pose estimation from RGB images are both highly challenging tasks due to the large discrepancy between factors of variation ranging from image background content to camera viewpoint. To better analyze these factors of variation, we propose the use of disentangled representations and a disentangled variational autoencoder (dVAE) that allows for specific sampling and inference of these factors. The derived objective from the variational lower bound as well as the proposed training strategy are highly flexible, allowing us to handle cross-modal encoders and decoders as well as semi-supervised learning scenarios. Experiments show that our dVAE can synthesize highly realistic images of the hand specifiable by both pose and image background content and also estimate 3D hand poses from RGB images with accuracy competitive with state-of-the-art on two public benchmarks.

1. Introduction

Vision-based hand pose estimation has progressed very rapidly in the past years [27, 38], driven in part by its potential for use in human-computer interaction applications. Advancements are largely due to the widespread availability of commodity depth sensors as well as the strong learning capabilities of deep neural networks. As a result, the majority of state-of-the-art methods apply deep learning methods to depth images [5, 6, 7, 8, 14, 18, 19, 32, 33]. Estimating 3D hand pose from single RGB images, however, is a less-studied and more difficult problem which has only recently gained some attention [3, 16, 21, 25, 40].

Unlike depth, which is a 2.5D source of information, RGB inputs have significantly more ambiguities. These ambiguities arise from the 3D to 2D projection and diverse backgrounds which are otherwise less pronounced in depth images. As such, methods which tackle the problem of monocular RGB hand pose estimation rely on learning from large datasets [40]. However, given the difficulties of accurately labelling hand poses in 3D, large-scale RGB datasets collected to date are synthesized [16, 40]. Real recorded datasets are much smaller, with only tens of sequences [30, 39]. This presents significant challenges when it comes to learning and motivates the need for strong kinematic and or image priors.

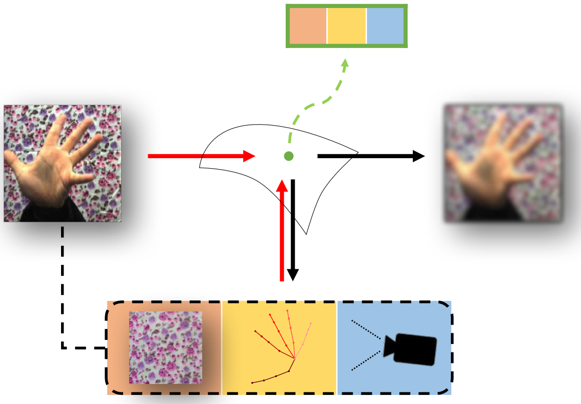


Figure 1: Illustration of dVAE. The red lines denote variational approximations while the black lines denote the generative model. With the help of labelled factors of variations (e.g. *pose, viewpoint and image content), we learn a disentangled and specifiable representation for RGB hand images in a VAE framework.*

Even though straight-forward discriminative approaches have shown great success in accurately estimating hand poses, there has also been growing interest in the use of deep generative models such as adversarial networks (GANs) [16, 32] and variational autoencoders (VAEs) [25]. Generative models can approximate and sample from the underlying distribution of hand poses as well as the associated images, and depending on the model formulation, may enable semi-supervised learning. This is particularly appealing for hand pose estimation, for which data with accurate ground truth can be difficult to obtain. One caveat, however, is that in their standard formulation, GANs and VAEs learn only black-box latent representations. Such representations offer little control for conditioning upon human-interpretable factors. Of the deep generative works presented to date [16, 25, 32], the latent representations are specifiable only by hand pose. Consequently it is possible to sample only a single (average) image per pose.

A recent work combining VAEs and GANs [4] introduced a conditional dependency structure to learn image backgrounds and demonstrated the possibility of transferring body poses onto different images. Inspired by this work, we would like to learn a similar latent representation that can disentangle the different factors that influence how hands may appear visually, *i.e.* normalized hand pose, camera viewpoint, scene context and background, *etc*. At the same time, we want to ensure that the disentangled representation remains sufficiently discriminative to make highly accurate estimates of 3D hand pose.

We present in this paper a disentangled variational autoencoder (dVAE) – a novel framework for learning disentangled representations of hand poses and hand images. As the factors that we would like to disentangle belong to different modalities, we begin with a cross-modal VAE [20, 25] as the baseline upon which we define our dVAE. By construction, our latent space is a disentangled one, composed of sub-spaces calculated by factors and a training strategy to fuse different latent space into one disentangled latent space. We show how these disentangled factors can be learned from both independent and confounding label inputs. To the best of our knowledge, our proposed model is the first disentangled representation that is able to both synthesize hand images and estimate hand poses with explicit control over the latent space. A schematic illustration of our dVAE and the disentangled factors is shown in Fig. 1. We summarize our contributions below:

- We propose a novel disentangled VAE model crossing different modalities; this model is the first VAE-based model that uses independent factors of variations to learn disentangled representations.
- Our dVAE model is highly flexible and handles multiple tasks including RGB hand image synthesis, pose transfer and 3D pose estimation from RGB images.
- We enable explicit control over different factors of variation and introduce the first model with multiple degrees of freedom for synthesizing hand images.
- We decouple the learning of disentangling factors and the embedding of image content and introduce two variants of learning algorithms for both independent and confounding labels.

2. Related Works

2.1. Hand Pose Estimation

Much of the progress made in hand pose estimation have focused on using depth image inputs [5, 6, 7, 8, 11, 14, 15, 18, 19, 32, 33, 35]. State-of-the-art methods use a convolutional neural network (CNN) architecture, with the majority of works treating the depth input as 2D pixels, though a few more recent approaches treat depth inputs as a set of 3D points and or voxels [7, 5, 15].

Estimating hand poses from monocular RGB inputs is more challenging. Early methods could recognize only a restricted set of poses [1, 36] or used simplified hand representations instead of full 3D skeletons [26, 37]. In more recent approaches, the use of deep learning and CNNs has become common-place [3, 21, 40]. In [16, 25], deep generative models such as variational auto-encoders (VAE) [25] and generative adversarial networks (GANs) [16] are applied, which makes feasible not only to estimate pose, but also generate RGB images from given hand poses.

Two hand pose estimation approaches [32, 25] stand out for being similar to ours in spirit. They also use shared latent spaces, even though the nature of these spaces are very different. Wan *et al.* [32] learns two separate latent spaces, one for hand poses and one for depth images, and uses a one-to-one mapping function to connect the two. Spurr *et al.* [25] learns a latent space that cross multiple hand modalities, such as RGB to pose and depth to pose. To force the cross-modality pairings onto a single latent space, separate VAEs are learned in an alternating fashion, with one input modality contributing to the loss per iteration. Such a learning strategy is non-ideal, as it tends to result in fluctuations in the latent space and has no guarantees for convergence. Additionally, by assuming all crossing modalities as one-to-one mappings, only one image can be synthesized per pose.

Different from [32] and [25], our dVAE learns a single latent space by design. We learn the latent space with the different modalities jointly, as opposed to alternating framework of [25]. We find that our joint learning is more stable and has better convergence properties. And because we explicitly model and disentangle image factors, we can handle one-to-many mappings, *i.e.* synthesize multiple images of the same hand pose.

2.2. Disentangled Representations

Disentangled representations separate data according to salient factors of variation and have recently been learned with deep generative models such as VAEs and GANs. Such representations have been applied successfully to image editing [2, 4, 13, 17, 24, 28], video generation [29] and image-to-image translation [12]. Several of these works [24, 28, 29, 34], however, require specially designed layers and loss functions, making the architectures difficult to work with and extend beyond their intended task.

Previous works learning disentangled representations with VAEs [2, 12, 13] typically require additional weak labels such as grouping information [2, 13] and pairwise similarities [12]. Such labels can be difficult to obtain and are often not defined for continuous variables such as hand pose and viewpoint. In [4, 17], a conditional dependency structure is proposed to train disentangled representations for a semi-supervised learning. The work of [4] resembles ours in the sense that they also disentangle pose from appearance; however, their conditional dependency structure is sensitive to the number of factors. As the number of factors grows, the complexity of the network structure increases exponentially. In comparison to existing VAE approaches, we are able to learn interpretable and disentangled representations by the shared latent space produced by image and its corresponding factors without additional weak labels.

3. Methodology

3.1. Cross Modal VAE

Before we present how a disentangled latent space can be incorporated into a VAE framework across different modalities, we first describe the original cross modal VAE [20, 25]. As the name suggests, the cross modal VAE aims to learn a VAE model across two different modalities $\mathbf{x}$ and $\mathbf{y}$. We begin by defining the log probability of the joint distribution $p(\mathbf{x},\mathbf{y})$. Since working with this distribution is intractable, one maximizes the evidence lower bound (ELBO) instead via a latent variable $\mathbf{z}$. Note that $\mathbf{x}$ and $\mathbf{y}$ are assumed to be conditionally independent given the latent $\mathbf{z}$, *i.e.* $(\mathbf{x} \perp \mathbf{y} \mid \mathbf{z})$.

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

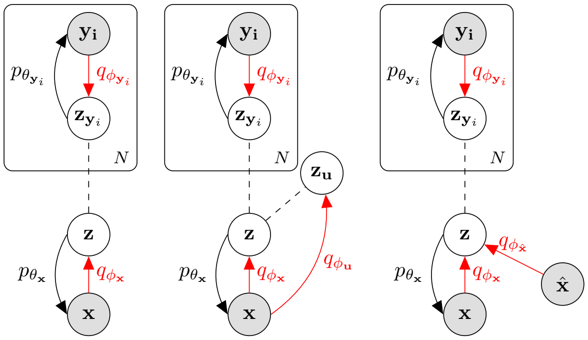
Here, $D_{KL}(\cdot)$ is the Kullback-Leibler divergence. The variational approximation $q_\phi(\mathbf{z}|\mathbf{x})$ can be thought of as an encoder from $\mathbf{x}$ to $\mathbf{z}$, while $p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z})$ and $p_{\theta_{\mathbf{y}}}(\mathbf{y}|\mathbf{z})$ can be thought of as decoders from $\mathbf{z}$ to $\mathbf{x}$ and $\mathbf{z}$ to $\mathbf{y}$ respectively. $p(\mathbf{z}) = \mathcal{N}(\mathbf{0},\mathbf{I})$ is a Gaussian prior on the latent space.

In the context of hand pose estimation, $\mathbf{x}$ would represent the RGB or depth image modality and $\mathbf{y}$ the hand skeleton modality. One can then estimate hand poses from images by encoding the image $\mathbf{x}$ into the latent space and decoding the corresponding 3D hand pose $\mathbf{y}$. A variant of this model was applied in [25] and shown to successfully estimate hand poses from RGB images or depth images.

3.2. Disentangled VAE

In our disentangled VAE, we define a latent variable $\mathbf{z}$ which can be deterministically decomposed into $N+1$ independent factors $\{\mathbf{z}_{\mathbf{y}_1}, \mathbf{z}_{\mathbf{y}_2}, ..., \mathbf{z}_{\mathbf{y}_N}, \mathbf{z}_{\mathbf{u}}\}$. Of these factors, $\{\mathbf{z}_{\mathbf{y}_i}\}_{i=1...N}$ are directly associated with observed variables $\{\mathbf{y}_i\}_{i=1...N}$. $\mathbf{z}_{\mathbf{u}}$ is an extra latent factor which is not independently associated with any observed variables; it may or may not be included (compare Fig. 2a versus Fig. 2b).

Fully specified latent $\mathbf{z}$: We begin first by considering the simplified case in which $\mathbf{z}$ can be fully specified by


$\mathbf{y_i}$
$p_{\theta_{\mathbf{y}_i}}$
$q_{\phi_{\mathbf{y}_i}}$
$\mathbf{z}_{\mathbf{y}_i}$
N
$\mathbf{z}$
$p_{\theta_{\mathbf{x}}}$
$q_{\phi_{\mathbf{x}}}$
$\mathbf{x}$
$\mathbf{z_u}$
$q_{\phi_{\mathbf{u}}}$
$q_{\phi_{\hat{\mathbf{x}}}}$
$\hat{\mathbf{x}}$

(a) dVAE (b) dVAE with $\mathbf{z_u}$ (c) dVAE with $\hat{\mathbf{x}}$

Figure 2: Graphical models of disentangled VAEs. The shaded nodes represent observed variables while un-shaded nodes are latent. The red and black solid lines denote variational approximations q_ϕ or encoders, and the generative models p_θ or decoders respectively. The dashed lines denote deterministically constructed variables. Figure best viewed in colour.

$\mathbf{z}_{\mathbf{y}_i}$ without $\mathbf{z_u}$, *i.e.* all latent factors can be associated with some observed $\mathbf{y}_i$. For clarity, we limit our explanation to $N=2$, though the theory generalizes to higher N as well. Our derivation can be separated into a disentangling step and an embedding step. In the ***disentangling step***, we first consider the joint distribution between $\mathbf{x}$, $\mathbf{y}_1$ and $\mathbf{y}_2$. The evidence lower bound of this distribution can be defined as:

$$\begin{aligned}\log p(\mathbf{x}, &\mathbf{y}_1,\mathbf{y}_2) \geq \text{ELBO}_{\text{dis}}(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2,\phi_{\mathbf{y}_1},\phi_{\mathbf{y}_2},\theta_{\mathbf{y}_1},\theta_{\mathbf{y}_2},\theta_{\mathbf{x}}) \\ &= \lambda_{\mathbf{x}} E_{\mathbf{z}\sim q_{\phi_{\mathbf{y}_1},\phi_{\mathbf{y}_2}}} \log p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z}) \\ &+ \lambda_{\mathbf{y}_1} E_{\mathbf{z}_{\mathbf{y}_1}\sim q_{\phi_{\mathbf{y}_1}}} \log p_{\theta_{\mathbf{y}_1}}(\mathbf{y}_1|\mathbf{z}_{\mathbf{y}_1}) \\ &+ \lambda_{\mathbf{y}_2} E_{\mathbf{z}_{\mathbf{y}_2}\sim q_{\phi_{\mathbf{y}_2}}} \log p_{\theta_{\mathbf{y}_2}}(\mathbf{y}_2|\mathbf{z}_{\mathbf{y}_2}) \\ &- \beta D_{KL}\left(q_{\phi_{\mathbf{y}_1},\phi_{\mathbf{y}_2}}(\mathbf{z}|\mathbf{y}_1,\mathbf{y}_2)||p(\mathbf{z})\right),\end{aligned} \tag{2}$$

where the λs and β are additional hyperparameters added to trade off between latent space capacity and reconstruction accuracy, as recommended by the β trick [10].

The ELBO in Eq. 2 allows us to define a disentangled $\mathbf{z} = [\mathbf{z}_{\mathbf{y}_1}, \mathbf{z}_{\mathbf{y}_2}]$ based on $\mathbf{y}_1$, $\mathbf{y}_2$ and $\mathbf{x}$. In this step, one can learn the encoding and decoding of $\mathbf{y}_i$ to and from $\mathbf{z}_{\mathbf{y}_i}$, as well as the decoding of $\mathbf{z}$ to $\mathbf{x}$. However, the mapping from $\mathbf{x}$ to $\mathbf{z}$ is still missing so we need an additional ***embedding step*** [31] to learn the encoder $q_{\phi_{\mathbf{x}}}(\mathbf{z}|\mathbf{x})$. Keeping all decoders fixed, $q_{\phi_{\mathbf{x}}}(\mathbf{z}|\mathbf{x})$ can be learned by maximizing:

$$\begin{aligned}\mathcal{L}(\phi_{\mathbf{x}}|\theta_{\mathbf{y}_1},\theta_{\mathbf{y}_2},\theta_{\mathbf{x}}) &= -D_{KL}\left(q_{\phi_{\mathbf{x}}}(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)\right) \\ &= \text{ELBO}_{\text{emb}}(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2,\phi_{\mathbf{x}}) - \log p(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2).\end{aligned} \tag{3}$$

Since the second term is constant with respect to $\phi_{\mathbf{x}}$ and the θ's, the objective simplifies to the following evidence lower bound with λ' and β' as hyperparameters:

$$\begin{aligned}\text{ELBO}_{\text{emb}}(\mathbf{x}{,}\mathbf{y}_1, \mathbf{y}_2, \phi_{\mathbf{x}}) = &\lambda'_{\mathbf{x}} E_{\mathbf{z}\sim q_{\phi_{\mathbf{x}}}} \log p_{\theta_{\mathbf{x}}}(\mathbf{x}|\mathbf{z}) \\ &+ \lambda'_{\mathbf{y}_1} E_{\mathbf{z}_{\mathbf{y}_1}\sim q_{\phi_{\mathbf{x}}}} \log p_{\theta_{\mathbf{y}_1}}(\mathbf{y}_1|\mathbf{z}_{\mathbf{y}_1}) \\ &+ \lambda'_{\mathbf{y}_2} E_{\mathbf{z}_{\mathbf{y}_2}\sim q_{\phi_{\mathbf{x}}}} \log p_{\theta_{\mathbf{y}_2}}(\mathbf{y}_2|\mathbf{z}_{\mathbf{y}_2}) \\ &- \beta' D_{KL}(q_{\phi_{\mathbf{x}}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})). \end{aligned} \tag{4}$$

Combining the disentangling and embedding evidence lower bounds, we get the following joint objective:

$$\begin{aligned}\mathcal{L}(\phi_{\mathbf{x}}{,}\phi_{\mathbf{y}_1}, \phi_{\mathbf{y}_2}, \theta_{\mathbf{x}}, \theta_{\mathbf{y}_1}, \theta_{\mathbf{y}_2}) = \\ \text{ELBO}_{\text{dis}}(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \phi_{\mathbf{y}_1}, \phi_{\mathbf{y}_2}, \theta_{\mathbf{x}}, \theta_{\mathbf{y}_1}, \theta_{\mathbf{y}_2}) \\ +\text{ELBO}_{\text{emb}}(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \phi_{\mathbf{x}}). \end{aligned} \tag{5}$$

The above derivation shows that the encoding of modality $\mathbf{x}$ can be decoupled from $\mathbf{y}_1$ and $\mathbf{y}_2$ via a disentangled latent space. We detail the training strategy for the fully specified version of the dVAE in Alg. 1.

Additional $\mathbf{z}_\mathbf{u}$: When learning a latent variable model, many latent factors may be very difficult to associate independently with an observation (label), *e.g.* the style of handwritten digits, or the background content in an RGB image [4, 13, 2]. Nevertheless, we may still want to disentangle such factors from those which can be associated independently. We model these factors in aggregate form via a single latent variable $\mathbf{z}_\mathbf{u}$ and show how $\mathbf{z}_\mathbf{u}$ can be disentangled from the other $\mathbf{z}_{\mathbf{y}_i}$ which are associated with direct observations $\mathbf{y}_i$. For clarity of discussion, we limit $N = 1$, such that $\mathbf{z} = [\mathbf{z}_{\mathbf{y}_1}, \mathbf{z}_\mathbf{u}]$. To disentangle $\mathbf{z}_\mathbf{u}$ from $\mathbf{z}$, both of which are specified by a confounding $\mathbf{x}$, we aim to make $\mathbf{z}_\mathbf{u}$ and $\mathbf{y}_1$ conditionally independent given $\mathbf{z}_{\mathbf{y}_1}$ To achieve this, we try to make $p(\mathbf{y}_1|\mathbf{z}_{\mathbf{y}_1}, \mathbf{z}_\mathbf{u})$ approximately equal to $p(\mathbf{y}_1|\mathbf{z}_{\mathbf{y}_1})$ and update the encoder and the decoder of $\mathbf{y}_1$ by random sampling of $\mathbf{z}_\mathbf{u}$ and minimizing the distance between $p(\mathbf{y}_1|\mathbf{z}_{\mathbf{y}_1}, \mathbf{z}_\mathbf{u})$ and $p(\mathbf{y}_1|\mathbf{z}_{\mathbf{y}_1})$. The training strategy for this is detailed in Alg. 2. In this case, the joint distribution of $\mathbf{x}$ and $\mathbf{y}_1$ has the following evidence lower bound in the ***disentangling step*** with hyperparameters λ'' and β'':

$$\begin{aligned}\log p(\mathbf{x}, \mathbf{y}_1) \geq\ &\text{ELBO}^{\mathbf{u}}_{\text{dis}}(\mathbf{x}, \mathbf{y}_1, \phi_{\mathbf{y}_1}, \phi_\mathbf{u}, \theta_{\mathbf{y}_1}, \theta_\mathbf{x}) \\ =&\lambda''_{\mathbf{x}} E_{\mathbf{z}\sim q_{\phi_{\mathbf{y}_1},\phi_\mathbf{u}}} \log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}) \\ +&\lambda''_{\mathbf{y}_1} E_{\mathbf{z}\sim q_{\phi_{\mathbf{y}_1},\phi_\mathbf{u}}} \log p_{\theta_{\mathbf{y}_1}}(\mathbf{y}_1|\mathbf{z}) \\ -&\beta'' D_{KL}(q_{\phi_{\mathbf{y}_1},\phi_\mathbf{u}}(\mathbf{z}|\mathbf{y}_1, \mathbf{x})||p(\mathbf{z})). \end{aligned} \tag{6}$$

Note that in the above ELBO, $\mathbf{z}_\mathbf{u}$ is encoded from $\mathbf{x}$ by $q_{\phi_\mathbf{u}}$ instead of being specified by some observed label $\mathbf{u}$, as was done previously in [13, 2, 4]. After this modified disentangling step, we can apply the same embedding step in Eq. 3 to learn $q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})$.

Algorithm 1 dVAE learning for fully specified $\mathbf{z}$.

```
Require: $\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \lambda_\mathbf{x}, \lambda_{\mathbf{y}_1}, \lambda_{\mathbf{y}_2}, \beta, T_1, T_2$
Ensure: $\phi_\mathbf{x}, \phi_{\mathbf{y}_1}, \phi_{\mathbf{y}_2}, \theta_\mathbf{x}, \theta_{\mathbf{y}_1}, \theta_{\mathbf{y}_2}$
 1: Initialize $\phi_\mathbf{x}, \phi_{\mathbf{y}_1}, \phi_{\mathbf{y}_2}, \theta_\mathbf{x}, \theta_{\mathbf{y}_1}, \theta_{\mathbf{y}_2}$
 2: for $t_1 = 1, \dots, T_1$ epochs do
 3:     Encode $\mathbf{y}_1, \mathbf{y}_2$ to $q_{\phi_{\mathbf{y}_1}}(\mathbf{z}_{\mathbf{y}_1}|\mathbf{y}_1), q_{\phi_{\mathbf{y}_2}}(\mathbf{z}_{\mathbf{y}_2}|\mathbf{y}_2)$
 4:     Construct $\mathbf{z} \leftarrow [\mathbf{z}_{\mathbf{y}_1}, \mathbf{z}_{\mathbf{y}_2}]$
 5:     Decode $\mathbf{z}$ to $p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}), p_{\theta_{\mathbf{y}_1}}(\mathbf{y}_1|\mathbf{z}_{\mathbf{y}_1}), p_{\theta_{\mathbf{y}_2}}(\mathbf{y}_2|\mathbf{z}_{\mathbf{y}_2})$
 6:     Update $\phi_{\mathbf{y}_1}, \phi_{\mathbf{y}_2}, \theta_{\mathbf{y}_1}, \theta_{\mathbf{y}_2}, \theta_\mathbf{x}$ via gradient ascent of
        Eq. 2
 7: end for
 8: for $t_2 = 1, \dots, T_2$ epochs do
 9:     Encode $\mathbf{x}$ to $q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})$
10:     Construct $[\mathbf{z}_{\mathbf{y}_1}, \mathbf{z}_{\mathbf{y}_2}] \leftarrow \mathbf{z}$
11:     Decode $\mathbf{z}$ to $p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}), p_{\theta_{\mathbf{y}_1}}(\mathbf{y}_1|\mathbf{z}_{\mathbf{y}_1}), p_{\theta_{\mathbf{y}_2}}(\mathbf{y}_2|\mathbf{z}_{\mathbf{y}_2})$
12:     Update $\phi_\mathbf{x}$ via gradient ascent of Eq. 4
13: end for
```

Multiple $\mathbf{x}$ modalities: The situation may arise in which we have multiple input modalities which fully specify and share the latent space of $\mathbf{z}$, *i.e.* not only an $\mathbf{x}$ but also an additional $\hat{\mathbf{x}}$ (see Fig. 2c). Here, it is possible to first consider the joint distribution between $\mathbf{x}$, $\mathbf{y}_1$ and $\mathbf{y}_2$, and maximize the ELBO in Eq. 2 for the disentangling step. To link the two modalities of $\mathbf{x}$ and $\hat{\mathbf{x}}$ into the same disentangled latent space and embed $\hat{\mathbf{x}}$, we can use the following:

$$\begin{aligned}\mathcal{L}(\phi_{\hat{\mathbf{x}}}|\theta_\mathbf{x}, \theta_{\mathbf{y}_1}, \theta_{\mathbf{y}_2}) &= -D_{KL}(q_{\phi_{\hat{\mathbf{x}}}}(\mathbf{z}|\hat{\mathbf{x}})||p_\theta(\mathbf{z}|\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)) \\ &= \text{ELBO}'_{\text{emb}}(\hat{\mathbf{x}}, \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \phi_{\hat{\mathbf{x}}}) - \log p(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2). \end{aligned} \tag{7}$$

Similar to Eq. 4, we get the following evidence lower bound with λ''' and β''' as hyperparameters:

$$\begin{aligned}\text{ELBO}'_{\text{emb}}(\hat{\mathbf{x}}, \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \phi_{\hat{\mathbf{x}}}) = &\lambda'''_{\mathbf{x}} E_{\mathbf{z}\sim q_{\phi_{\hat{\mathbf{x}}}}} \log p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}) \\ &+ \lambda'''_{\mathbf{y}_1} E_{\mathbf{z}_{\mathbf{y}_1}\sim q_{\phi_{\hat{\mathbf{x}}}}} \log p_{\theta_{\mathbf{y}_1}}(\mathbf{y}_1|\mathbf{z}_{\mathbf{y}_1}) \\ &+ \lambda'''_{\mathbf{y}_2} E_{\mathbf{z}_{\mathbf{y}_2}\sim q_{\phi_{\hat{\mathbf{x}}}}} \log p_{\theta_{\mathbf{y}_2}}(\mathbf{y}_2|\mathbf{z}_{\mathbf{y}_2}) \\ &- \beta''' D_{KL}(q_{\phi_{\hat{\mathbf{x}}}}(\mathbf{z}|\hat{\mathbf{x}})||p(\mathbf{z})). \end{aligned} \tag{8}$$

For learning, one simply encodes $\hat{\mathbf{x}}$ with $q_{\phi_{\hat{\mathbf{x}}}}(\mathbf{z}|\hat{\mathbf{x}})$ to $\mathbf{z}$ instead of $p_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})$ as shown currently in line 9 of Alg. 1. A full derivation of the dVAE and its variants is given in the supplementary.

3.3. Applications

Based on the theory proposed above, we develop two applications: image synthesis and pose estimation from RGB images. Like [40], we distinguish between an absolute 3D hand pose (3DPose), a canonical hand pose (CPose), and a viewpoint. The canonical pose is a normalized version of the 3D pose within the canonical frame, while viewpoint is the rotation matrix that rotates CPose to 3DPose.

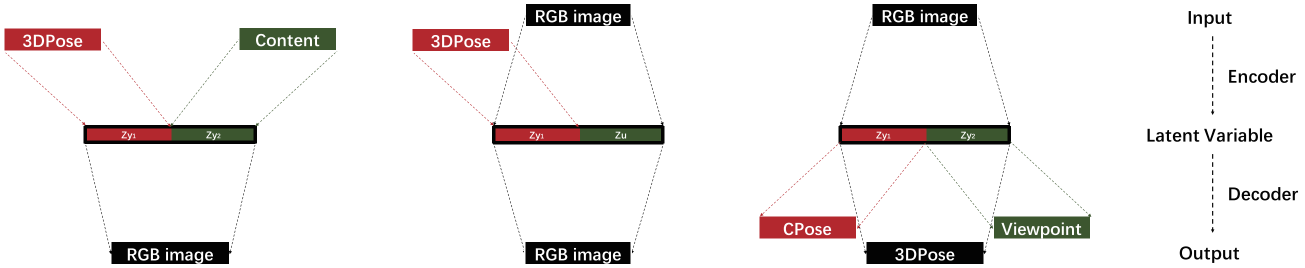


Figure 3: Inference models for the tasks of image synthesis (left and middle) and pose estimation (right).

In **image synthesis**, we would like to sample values of $\mathbf{z}$ and decode this into an image $\mathbf{x}$ via the generative model $p_{\theta_{\mathbf{x}}}$. To control the images being sampled, we want to have a latent $\mathbf{z}$ which is disentangled with respect to the 3DPose, and image (background) content, *i.e.* all aspects of the RGB image not specifically related to the hand pose itself. A schematic of the image synthesis is shown in the left panel of Fig. 3; in this case, we follow the model in Fig. 2a and use Alg. 1. Here, $\mathbf{y}_1$ would represent 3DPose and $\mathbf{y}_2$ would represent the image content; similar to [29], this content is specified by a representative tag image. By changing the inputs $\mathbf{y}_1$ and $\mathbf{y}_2$, *i.e.* by varying the 3DPose and content through the encoders $q_{\phi_{\mathbf{y}_1}}$ and $q_{\phi_{\mathbf{y}_2}}$, we synthesize new images with specified poses and background content. Furthermore, we can also evaluate the pose error of the synthesized image via the pose decoder $p_{\theta_{\mathbf{y}_1}}$.

Tag images for specifying background content are easy to obtain if one has video sequences from which to extract RGB frames. However, for some scenarios, this may not be the case, *i.e.* if each RGB image in the training set contains different background content. This is what necessitates the model in Fig. 2b and the learning algorithm in Alg. 2. In such a scenario, $\mathbf{y}_1$ again represents the 3DPose, while the image content is modelled indirectly through $\mathbf{x}$. For testing purposes, however, there is no distinction between the two variants, as input is still given in the form of a desired 3DPose and an RGB image specifying the content.

Algorithm 2 dVAE learning for additional $\mathbf{z}_\mathbf{u}$.

```
Require: $\mathbf{x}, \mathbf{y}_1, \lambda_\mathbf{x}, \lambda_{\mathbf{y}_1}, \beta, T_1, T_2, T_3$
Ensure: $\phi_\mathbf{x}, \phi_{\mathbf{y}_1}, \phi_\mathbf{u}, \theta_\mathbf{x}, \theta_{\mathbf{y}_1}$
Initialize $\phi_\mathbf{x}, \phi_{\mathbf{y}_1}, \phi_\mathbf{u}, \theta_\mathbf{x}, \theta_{\mathbf{y}_1}$
for $t_1 = 1, \ldots, T_1$ epochs do
    Encode $\mathbf{x}, \mathbf{y}_1$ to $q_{\phi_{\mathbf{y}_1}}(\mathbf{z}_{\mathbf{y}_1}|\mathbf{y}_1), q_{\phi_\mathbf{u}}(\mathbf{z}_\mathbf{u}|\mathbf{x})$
    Construct $\mathbf{z} \leftarrow [\mathbf{z}_{\mathbf{y}_1}, \mathbf{z}_\mathbf{u}]$, $[\mu, \sigma] \leftarrow q_{\phi_\mathbf{u}}(\mathbf{z}_\mathbf{u}|\mathbf{x})$
    Decode $\mathbf{z}$ to $p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}), p_{\theta_{\mathbf{y}_1}}(\mathbf{y}_1|\mathbf{z})$
    Update $\phi_{\mathbf{y}_1}, \phi_\mathbf{u}, \theta_{\mathbf{y}_1}, \theta_\mathbf{x}$
    for $t_2 = 1, \ldots, T_2$ epochs do
        Encode $\mathbf{y}_1$ to $q_{\phi_{\mathbf{y}_1}}(\mathbf{z}_{\mathbf{y}_1}|\mathbf{y}_1)$
        Construct $\mathbf{z}_{noise} \leftarrow \mathcal{N}(\mu, \sigma)$, $\mathbf{z} \leftarrow [\mathbf{z}_{\mathbf{y}_1}, \mathbf{z}_{noise}]$
        Decode $\mathbf{z}$ to $p_{\theta_{\mathbf{y}_1}}(\mathbf{y}_1|\mathbf{z})$
        Update $\phi_{\mathbf{y}_1}, \theta_{\mathbf{y}_1}$
    end for
end for
for $t_3 = 1, \ldots, T_3$ epochs do
    Encode $\mathbf{x}$ to $q_{\phi_\mathbf{x}}(\mathbf{z}|\mathbf{x})$
    Construct $[\mathbf{z}_{\mathbf{y}_1}, \mathbf{z}_\mathbf{u}] \leftarrow \mathbf{z}$
    Decode $\mathbf{z}$ to $p_{\theta_\mathbf{x}}(\mathbf{x}|\mathbf{z}), p_{\theta_{\mathbf{y}_1}}(\mathbf{y}_1|\mathbf{z})$
    Update $\phi_\mathbf{x}$
end for
```

For **hand pose estimation**, we aim to predict 3DPose $\mathbf{x}$, CPose $\mathbf{y}_1$ and viewpoint $\mathbf{y}_2$ from RGB image $\hat{\mathbf{x}}$ according to the model in Fig. 2c by disentangling $\mathbf{z}$ into the CPose $\mathbf{z}_{\mathbf{y}_1}$ and viewpoint $\mathbf{z}_{\mathbf{y}_2}$. In this case, we embed $\mathbf{x}$ and $\hat{\mathbf{x}}$ into a shared latent space. We apply inference as shown by the right panel in Fig. 3 and learn the model with Alg. 1. Moreover, because annotated training data is sparse in real world applications, we can further leverage unlabelled or weakly labelled. Our proposed method consists of multiple VAEs, which can be trained respectively for semi- and weakly-supervised setting. For semi-supervised setting, we use both labelled and unlabelled CPose, viewpoint and 3DPose data to train the encoders $q_{\phi_{\mathbf{y}_1}}$,$q_{\phi_{\mathbf{y}_2}}$ and all decoders in the disentangled step. For weakly-supervised setting, we exploit images and their weak labels like viewpoint $\mathbf{y}_2$ by training the VAE with $q_{\phi_{\hat{\mathbf{x}}}}$ and $p_{\theta_{\mathbf{y}_2}}$ in the embedding step.

4. Experimentation

A good disentangled representation should show good performance on both discriminative tasks such as hand pose estimation as well as generative tasks. We transfer attributes between images and infer 3D hand poses from monocular hand RGB images via disentangled representations. More precisely, for image synthesis, we transfer image content with fixed 3DPose, while for 3D hand pose estimation, we predict viewpoint, CPose and 3DPose.

4.1. Implementation details

Our architecture consists of multiple encoders and decoders. For encoding images, we use Resnet-18 [9]; for

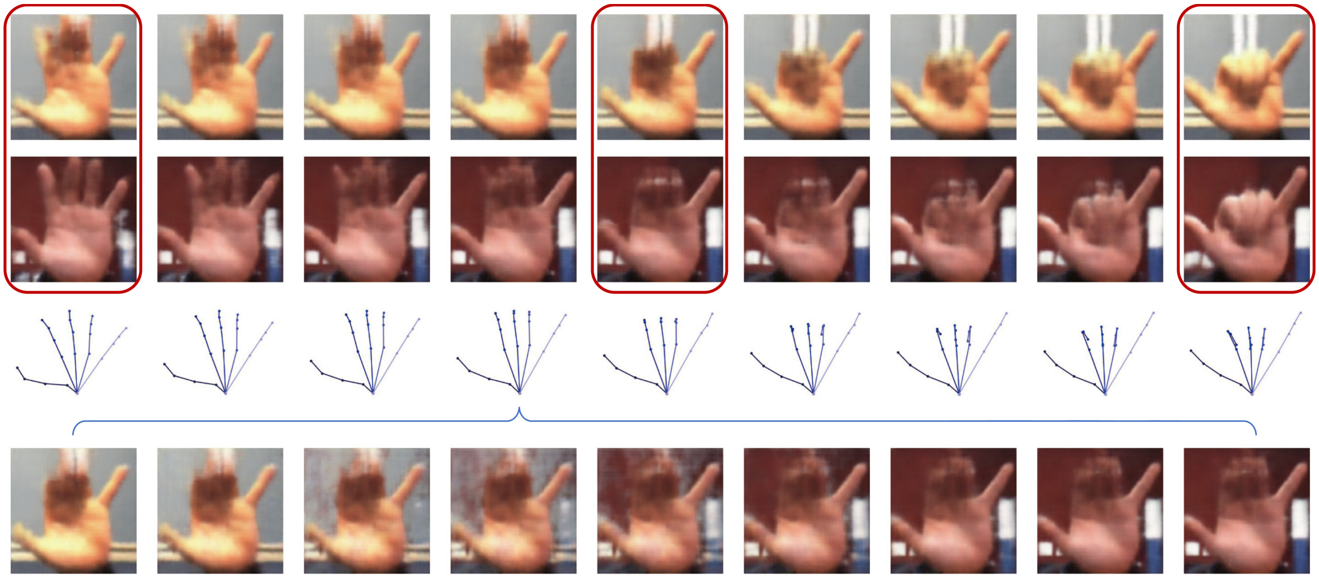


Figure 4: Latent space walk. The images in the red boxes are provided inputs. The first two rows show synthesized images when interpolating on the latent 3DPose space; the third row shows skeletons of the reconstructed 3DPose. The fourth row shows synthesized images when the pose is fixed (to the fourth column) when interpolating in the content latent space.

decoding images, we follow the decoder architecture DC-GAN [22]. For encoding and decoding hand poses, we use six fully connected layers with 512 hidden units. Exact architectural specifications are provided in the supplementary.

For learning, we use the ADAM optimizer with a learning rate of 10^{-4}, a batch size of 32. We fix the dimensionality of d of $\mathbf{z}$ to 64 and set the dimensionality of sub-latent variable $\mathbf{z}_{\mathbf{y}_1}$ and $\mathbf{z}_{\mathbf{y}_2}$ to 32 and 32. For all applications, the λ's are fixed ($\lambda_{\mathbf{x}} = 1, \lambda_{\mathbf{y}_1} = \lambda_{\mathbf{y}_2} = 0.01$) while we must adjust β ($\beta = 100$ for image synthesis, $\beta''' = 0.01$ for pose estimation). Further discussion on the impact of β and d can be found in the supplementary.

4.2. Datasets & Evaluation

We evaluate our proposed method on two publicly available datasets: Stereo Hand Pose Tracking Benchmark (STB) [39] and Rendered Hand Pose Dataset (RHD) [40].

The **STB dataset** features videos of a single person's left hand in front of 6 real-world indoor backgrounds. It provides the 3D positions of palm and finger joints for approximately 18k stereo pairs with 640 × 480 resolution. Image synthesis is relatively easy for this dataset due to the small number of backgrounds. To evaluate our model's pose estimation accuracy, we use the 15k / 3k training/test split as given by [40]. For evaluating our dVAE's generative modelling capabilities, we disentangle $\mathbf{z}$ into two content and 3DPose according to the model in Fig. 2a synthesize images with fixed poses as per the left-most model in Fig. 3.

RHD is a synthesized dataset of rendered hand images with 320×320 resolution from 20 characters performing 39 actions with various hand sizes, viewpoints and backgrounds. The dataset is highly challenging due to the diverse visual scenery, illumination and noise. It is composed of 42k images for training and 2.7k images for testing.

For quantitative evaluation and comparison with other works on 3D hand pose estimation, we use the common metrics, mean end-point-error (EPE) and the area under the curve (AUC) on the percentage of correct keypoints (PCK) score. Mean EPE is defined as the average euclidean distance between predicted and groundtruth keypoints; PCK is the percentage of predicted keypoints that fall within some given distance with respect to the ground truth.

4.3. Synthesizing Images

We evaluate the ability of our model to synthesize images by sampling from latent space walks and by transferring pose from one image to another.

For the **fully specified latent** $\mathbf{z}$ model we show the synthesized images (see Fig. 4) when we interpolate the 3DPose while keeping the image content fixed (rows 1-3) and when we interpolate image content while keeping the pose fixed. In both latent space walks, the reconstructed poses as well as the synthesized images demonstrate a smoothness and consistency of the latent space.

We can also extract disentangled latent factors from different hand images and then recombine them to transfer poses from one image to another. Fig. 6 shows the results when we take poses from one image (leftmost column), content from other images (top row) and recombine them (rows 2-3, columns 3-5). We are able to accurately transfer

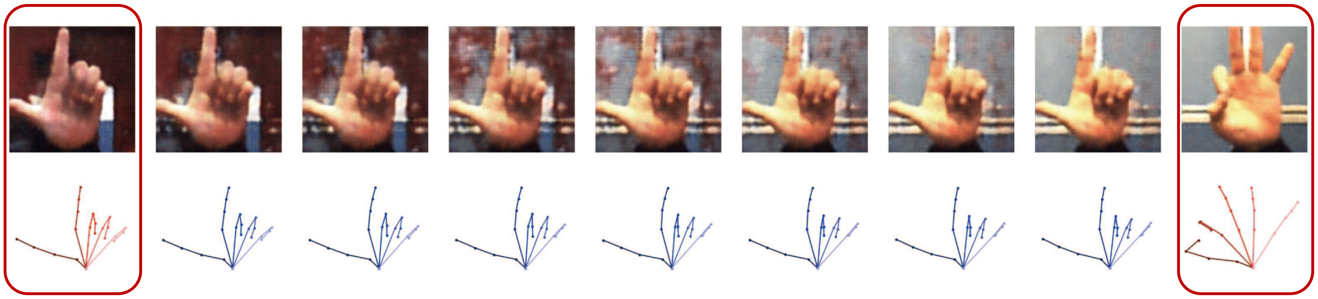


Figure 5: Latent space walk, interpolating $\mathbf{z_u}$ representing image background content. The images along with groundtruth 3DPose (red) in the red box are the input points; the first row shows generated images and the second row corresponding reconstructed 3DPose (blue). Note that because we are interpolating only on the background content, the pose stays well-fixed.

the hand poses while faithfully maintaining the tag content.

With additional $\mathbf{z_u}$ we also show interpolated results from a latent space walk on $\mathbf{z_u}$ in Fig. 5. In this case, the 3DPose stays well-fixed, while the content changes smoothly between the two input images, demonstrating our model's ability to disentangle the image background content even with out specific tag images for training.

4.4. 3D hand pose estimation

We evaluate the ability of our dVAE to estimate 3D hand poses from RGB images based on the model variant described in Section 3.3 and compare against state-of-the-art methods [3, 25, 40, 16, 21] on both the RHD and STB datasets. In [40], a two-stream architecture is applied to estimate viewpoint and CPose; these two are then combined to predict 3DPose. To be directly comparable, we disentangle the latent $\mathbf{z}$ into a viewpoint factor and a CPose factor, as shown in Fig. 3 right. Note that due to the decompositional nature of our latent space, we can predict viewpoint, CPose and 3DPose through one latent space.

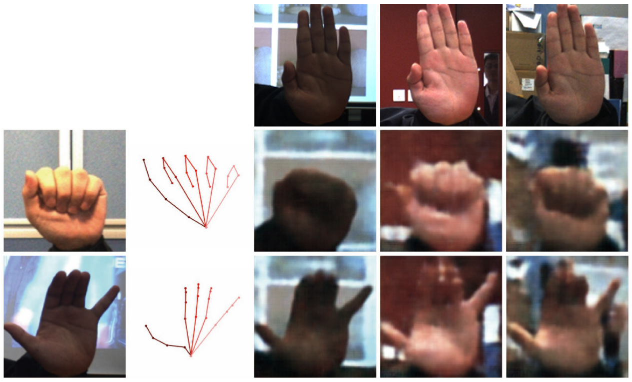


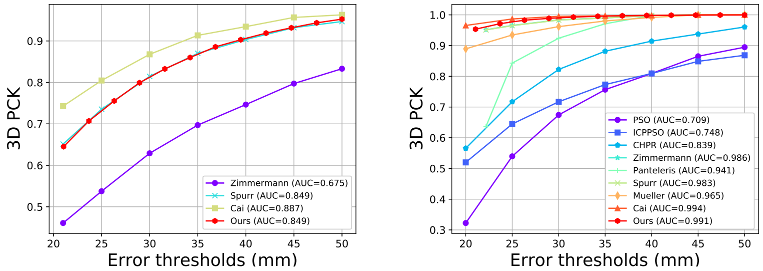
Figure 6: Pose transfer. The first column corresponds to images from which we extract the 3DPose (ground truth pose in second column); the first row corresponds to tag images columns we extract the latent content; the 2-3 rows, 3-5 columns are pose transferred images.

We follow the experimental setting in [40, 25] that left vs right handedness and scale are given at test time. We augment the training data by rotating the images in the range of $[-180°, 180°]$ and making random flips along the y-axis while applying the same transformations to the ground truth labels. We compare the mean EPE in Fig. 7 right. We outperform [40] on both CPose and 3DPose. These results highlight the strong capabilities of our dVAE model for accurate hand pose estimation. Our mean EPE is very close to that of [25], while our 3D PCK is slightly better. As such, we conclude that the pose estimation capabilities of our model is comparable to that of [25], though our model is able to obtain a disentangled representation and make full use of weak labels. We compare the PCK curves with state-of-the-art methods [3, 25, 40, 16, 21] on both datasets in Fig. 7. Our method is comparable or better than most existing methods except [3], which has a higher AUC of 0.038 on RHD and 0.03 on STB for the PCK. However, these results are not directly comparable, as [3] incorporate depth images as an additional source of training data. Fig. 8 shows some our estimated hand poses from both RHD and STB datasets.

Semi-, weakly-supervised learning: To evaluate our method in semi- and weakly-supervised settings, we sample the first $m\%$ images as labelled data and the rest as unlabelled data by discarding the labels of 3DPose, CPose and viewpoint. We also consider using only viewpoints as a weak label while discarding 3DPose and CPose. For the RHD dataset, we vary $m\%$ from 5% to 100% and compare the mean EPE against the fully supervised setting. We can see that our model makes full use of additional information. With CPose, viewpoint and 3DPose labels, we improve the mean EPE up to 3.5%. With additional images and viewpoint labels, the improvement is up to 7.5%.

5. Conclusion

In this paper, we presented a VAE-based method for learning disentangled representations of hand poses and hand images. We find that our model allows us to synthe-



Method	RHD		STB	
	CPose	3DPose	CPose	3DPose
[40]	16.37	30.42	6.07	8.68
[25]	\	19.73	\	8.56
Ours	13.93	19.95	6.09	8.66

Figure 7: Quantitative evaluation. 3D PCK on RHD (left) and STB (middle). Mean EPE (mm) on RHD and STB (right).

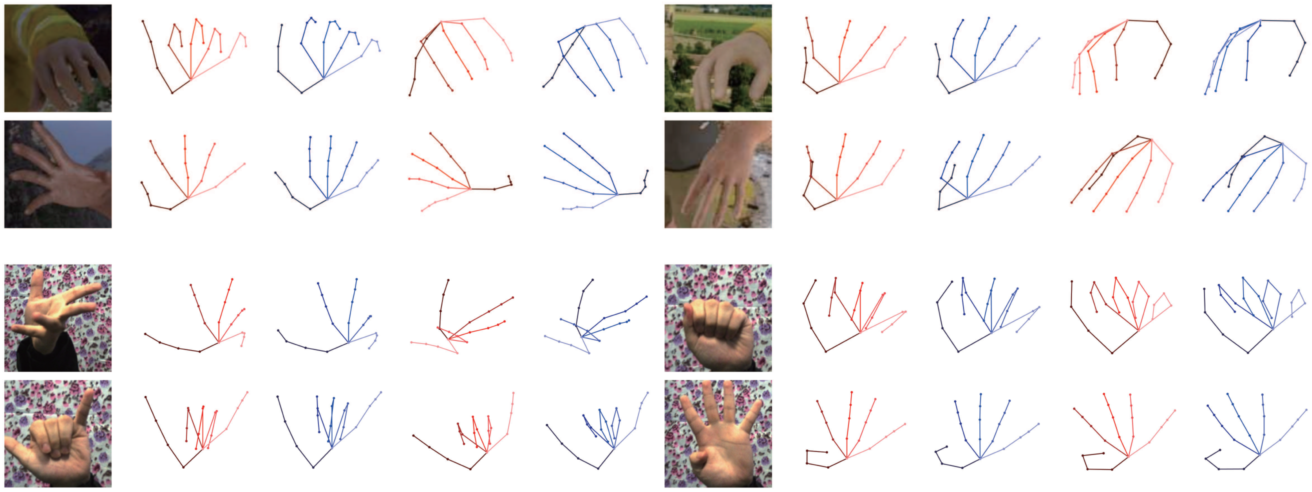


Figure 8: CPose and 3DPose estimation on RHD and STB. For each quintet, the left most column corresponds to the input images, the second and the third columns correspond to CPose groundtruth (red) and our prediction (blue), the right most two columns correspond to 3DPose groundtruth (red) and our prediction (blue).

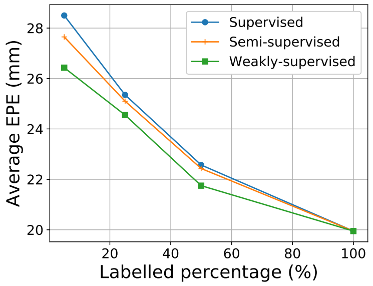


Figure 9: Mean EPE of our model on the semi-supervised setting and the weakly-supervised setting.

size highly realistic looking RGB images of hands with full control over factors of variation such as image background content and hand pose. However, the factors of variation here should be independent. This is a valid assumption for hand images, but we will consider to relax the need of independence between factors and further investigate disentangled representations with multimodal learning.

For hand pose estimation, our model is competitive with state of the art and is also able to leverage unlabelled and weak labels. Currently, STB is the standard benchmark for real-world monocular RGB hand pose estimation. However, since the featured background content and hand poses are quite simple, performance by state-of-the-art methods on this dataset has become saturated. For the 3D PCK, recent works [3, 25, 40, 16, 21] achieve AUC values for error thresholds of 20-50mm ranging from 96% to more than 99%. As such, we encourage members of the community to collect more challenging benchmarks for RGB hand pose estimation. In particular, for the monocular scenario, one possibility would be to collect multi-view [23] and also multi-modal data, *i.e.* RGBD, from which it is possible to use highly accurate model-based trackers to estimate ground truth labels.

Acknowledgments Research in this paper was partly supported by the Singapore Ministry of Education Academic Research Fund Tier 1.

References

[1] Vassilis Athitsos and Stan Sclaroff. Estimating 3d hand pose from a cluttered image. In *CVPR*. IEEE, 2003. 2

[2] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI*, 2018. 2, 4

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

[4] Rodrigo de Bem, Arnab Ghosh, Thalaiyasingam Ajanthan, Ondrej Miksik, N Siddharth, and Philip HS Torr. A semi-supervised deep generative model for human body analysis. In *ECCVW*, 2018. 2, 3, 4

[5] Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *CVPR*, 2018. 1, 2

[6] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3D hand pose estimation in single depth images: from Single-View CNN to Multi-View CNNs. In *CVPR*, 2016. 1, 2

[7] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *CVPR*, 2017. 1, 2

[8] Hengkai Guo, Guijin Wang, Xinghao Chen, Cairong Zhang, Fei Qiao, and Huazhong Yang. Region ensemble network: Improving convolutional network for hand pose estimation. In *ICIP*, 2017. 1, 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[10] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2016. 3

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

[12] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *ECCV*, 2018. 2

[13] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015. 2, 4

[14] Meysam Madadi, Sergio Escalera, Alex Carruesco, Carlos Andujar, Xavier Baró, and Jordi Gonzàlez. Occlusion aware hand pose recovery from sequences of depth images. In *FG*, 2017. 1, 2

[15] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *CVPR*, 2018. 2

[16] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3D hand tracking from monocular RGB. In *CVPR*, 2018. 1, 2, 7, 8

[17] Siddharth Narayanaswamy, T Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In *NIPS*, 2017. 2

[18] Markus Oberweger and Vincent Lepetit. Deepprior++: Improving fast and accurate 3D hand pose estimation. In *ICCVW*, 2017. 1, 2

[19] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. In *CVWW*, 2015. 1, 2

[20] Gaurav Pandey and Ambedkar Dukkipati. Variational methods for conditional multimodal deep learning. In *IJCNN*, 2017. 2, 3

[21] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single RGB frame for real time 3D hand pose estimation in the wild. In *WACV*, 2018. 1, 2, 7, 8

[22] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *arXiv preprint arXiv:1511.06434*, 2015. 6

[23] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *CVPR*, pages 8437–8446, 2018. 8

[24] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 2

[25] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 1, 2, 3, 7, 8

[26] Bjoern Stenger, Paulo RS Mendonça, and Roberto Cipolla. Model-based 3d tracking of an articulated hand. In *CVPR*. IEEE, 2001. 2

[27] James S Supancic, Gregory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan. Depth-Based hand pose estimation: Data, methods, and challenges. In *ICCV*, 2015. 1

[28] Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. In *arXiv preprint arXiv:1711.02245*, 2017. 2

[29] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018. 2, 5

[30] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 118, 2016. 1

[31] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. In *ICLR*, 2018. 3

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

[33] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In *CVPR*, 2018. 1, 2

[34] Chaoyue Wang, Chaohui Wang, Chang Xu, and Dacheng Tao. Tag disentangled generative adversarial networks for object image re-rendering. In *IJCAI*, 2017. 2

[35] Jan Wöhlke, Shile Li, and Dongheui Lee. Model-based hand pose estimation for generalized hand shape with appearance normalization. In *arXiv preprint arXiv:1807.00898*, 2018. 2

[36] Ying Wu and Thomas S Huang. View-independent recognition of hand postures. In *CVPR*. IEEE, 2000. 2

[37] Ying Wu, John Y Lin, and Thomas S Huang. Capturing natural hand articulation. In *ICCV*, page 426. IEEE, 2001. 2

[38] Shanxin Yuan, Guillermo Garcia-Hernando, Bjorn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. 3d hand pose estimation: From current achievements to future goals. In *CVPR*, 2018. 1

[39] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3D hand pose tracking and estimation using stereo matching. In *arXiv preprint arXiv:1610.07214*, 2016. 1, 6

[40] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, 2017. 1, 2, 4, 6, 7, 8