Cross-modal Deep Variational Hand Pose Estimation

Adrian Spurr, Jie Song, Seonwook Park, Otmar Hilliges
ETH Zurich
{spurra,jsong,spark,otmarh}@inf.ethz.ch

Abstract

The human hand moves in complex and high-dimensional ways, making estimation of 3D hand pose configurations from images alone a challenging task. In this work we propose a method to learn a statistical hand model represented by a cross-modal trained latent space via a generative deep neural network. We derive an objective function from the variational lower bound of the VAE framework and jointly optimize the resulting cross-modal KL-divergence and the posterior reconstruction objective, naturally admitting a training regime that leads to a coherent latent space across multiple modalities such as RGB images, 2D keypoint detections or 3D hand configurations. Additionally, it grants a straightforward way of using semi-supervision. This latent space can be directly used to estimate 3D hand poses from RGB images, outperforming the state-of-the art in different settings. Furthermore, we show that our proposed method can be used without changes on depth images and performs comparably to specialized methods. Finally, the model is fully generative and can synthesize consistent pairs of hand configurations across modalities. We evaluate our method on both RGB and depth datasets and analyze the latent space qualitatively.

1. Introduction

Hands are of central importance to humans in manipulating the physical world and in communicating with each other. Recovering the spatial configuration of hands from natural images therefore has many important applications in AR/VR, robotics, rehabilitation and HCI. Much work exists that tracks articulated hands in streams of depth images, or that estimates hand pose [15, 16, 27, 35] from individual depth frames. However, estimating the full 3D hand pose from monocular RGB images only is a more challenging task due to the manual dexterity, symmetries and self-similarities of human hands as well as difficulties stemming from occlusions, varying lighting conditions and lack of accurate scale estimates. Compared to depth images the RGB case is less well studied.

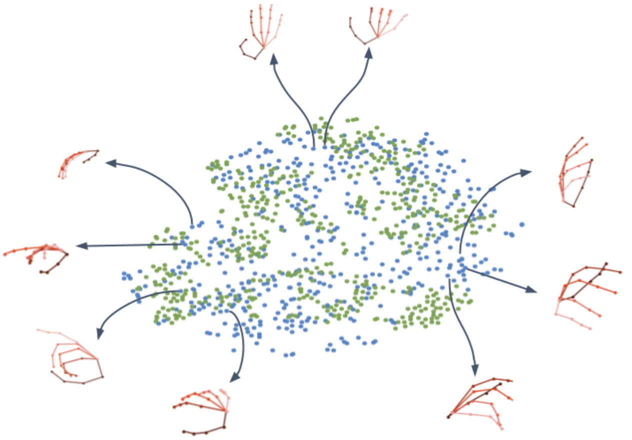


Figure 1: **Cross-modal latent space.** t-SNE visualization of 500 input samples of different modalities in the latent space. Embeddings of RGB images are shown in blue, embeddings of 3D joint configurations in green. Hand poses are decoded samples drawn from the latent space. Embedding does not cluster by modality, showing that there is a unified latent space. The posterior across different modalities can be estimated by sampling from this manifold.

Recent work relying solely on RGB images [38] proposes a deep learning architecture that decomposes the task into several substeps, demonstrating initial feasibility and providing a public dataset for comparison. The proposed architecture is specifically designed for the monocular case and splits the task into hand and 2D keypoint detection followed by a 2D-3D lifting step but incorporates no explicit hand model. Our work is also concerned with the estimation of 3D joint-angle configurations of human hands from RGB images but learns a cross-modal, statistical hand model. This is attained via learning of a latent representation that embeds sample points from multiple data sources such as 2D keypoints, images and 3D hand poses. Samples from this latent space can then be reconstructed by *independent* decoders to produce *consistent* and physically plausible 2D or 3D joint predictions and even RGB images.

Findings from bio-mechanics suggest that while articulated hands have many degrees-of-freedom, only few are fully independently articulated [20]. Therefore a sub-space of valid hand poses is supposed to exist and prior work on depth based hand tracking [26] has successfully employed dimensionality reduction techniques to improve accuracy.

This idea has been recently revisited in the context of deep-learning, where Wan et al. [34] attempt to learn a manifold of hand poses via a combination of variational autoencoders (VAEs) and generative adversarial networks (GANs) for hand pose estimation from depth images. However, their approach is based on two separate manifolds, one for 3D hand joints (VAE) and one for depth-maps (GAN) and requires a mapping function between the two.

In this work we propose to learn a single, unified latent space via an extension of the VAE framework. We provide a derivation of the variational lower bound that permits training of a single latent space using multiple modalities, where similar input poses are embedded close to each other independent of the input modality. Fig. 1 visualizes this learned unified latent space for two modalities (RGB & 3D). We focus on RGB images and hence test the architecture on different combinations of modalities where the goal is to produce 3D hand poses as output. At the same time, the VAE framework naturally allows to generate samples consistently in any modality.

We experimentally show that the proposed approach outperforms the state-of-the art method [38] in direct RGB to 3D hand pose estimation, as well as in lifting from 2D detections to 3D on a challenging public dataset. Meantime, we note that given any input modality a mapping into the embedding space can be found and likewise hand configurations can be reconstructed in various modalities, thus the approach learns a many-to-many mapping. We demonstrate this capability via generation of novel hand pose configurations via sampling from the latent space and consistent reconstruction in different modalities (i.e., 3D joint positions and synthesized RGB images). These could be potentially used in hybrid approaches for temporal tracking or to generate additional training data. Furthermore, we explore the utility of the same architecture in the case of depth images and show that we are comparable to state-of-art depth based methods [15, 16, 34] that employ specialized architectures.

2. Related Work

Capturing the 3D motion of human hands from images is a long standing problem in computer vision and related areas (cf. [5]). With the recent emergence of consumer grade RGB-D sensors and increased importance of AR and VR this problem has seen increased attention [22, 25, 26, 27, 28, 29, 30, 34, 35, 37]. Generally speaking approaches can be categorized into tracking of articulated hand motion over time (e.g., [18]) and per-frame classification [25, 27, 34]. Furthermore, a number of hybrid methods exist that first leverage a discriminative model to initialize a hand pose estimate which is then refined and tracked via carefully designed energy functions to fit a hand model into the observed depth data [19, 22, 30, 33, 36]. Estimating hand pose from RGB images is more challenging.

Also using depth-images, a number of approaches have been proposed that extract manually designed features and discriminative machine learning models to predict joint locations in depth images or 3D joint-angles directly [3, 10, 25, 28]. More recently a number of deep-learning models have been proposed that take depth images as input and regress 2D joint locations in multiple images [24, 32] which are then used for optimization-based hand pose estimation. Others deploy convolutional neural networks (CNNs) in end-to-end learning frameworks to regress 3D hand poses from depth images, either directly estimating 3D joint configurations [15, 23], or estimating joint-angles instead of Cartesian coordinates [16]. Exploiting the depth information more directly, it has also been proposed to convert depth images into 3D multi-views [6] or volumetric representations [7] before feeding them to a 3D CNN. Aiming at more mobile usage scenarios, recent work has proposed hybrid methods for hand-pose estimation from body-worn cameras under heavy occlusion [13]. While the main focus lies on RGB imagery, our work is also capable of predicting hand pose configurations from depth images due to the multi-modal latent space.

Wan et al. [34] is the most related work in spirit to ours. Like our work, they employ deep generative models (a combination of VAEs and GANs) to learn a latent space representation that regularizes the posterior prediction. Our method differs significantly in that we propose a theoretically grounded derivation of a cross-modal training scheme based on the variational autoencoder [11] framework that allows for joint training of a single cross-modal latent space, whereas [34] requires training of two separate latent spaces, learning of a mapping function linking them and final end-to-end refinement. Furthermore, we experimentally show that our approach reaches parity with the state-of-the-art in depth based hand pose estimation *and* outperforms existing methods in the RGB case, whereas [34] report only depth based experiments. In [2], VAE is also deployed for depth based hand pose estimation. However, their focus is minimising the dissimilarity coefficient between the true distribution and the estimated distribution.

To the best of our knowledge there is currently only one approach for learning-based hand pose estimation from RGB images alone [38]. Demonstrating the feasibility of the task, this work splits 3D hand pose estimation into an image segmentation, 2D joint detection and 2D-3D lifting task. Our approach allows for training of the latent space using either input modality (in this case 2D key points or RGB images) and direct 3D hand pose estimation via decoding the corresponding sample from the latent space. We exper-

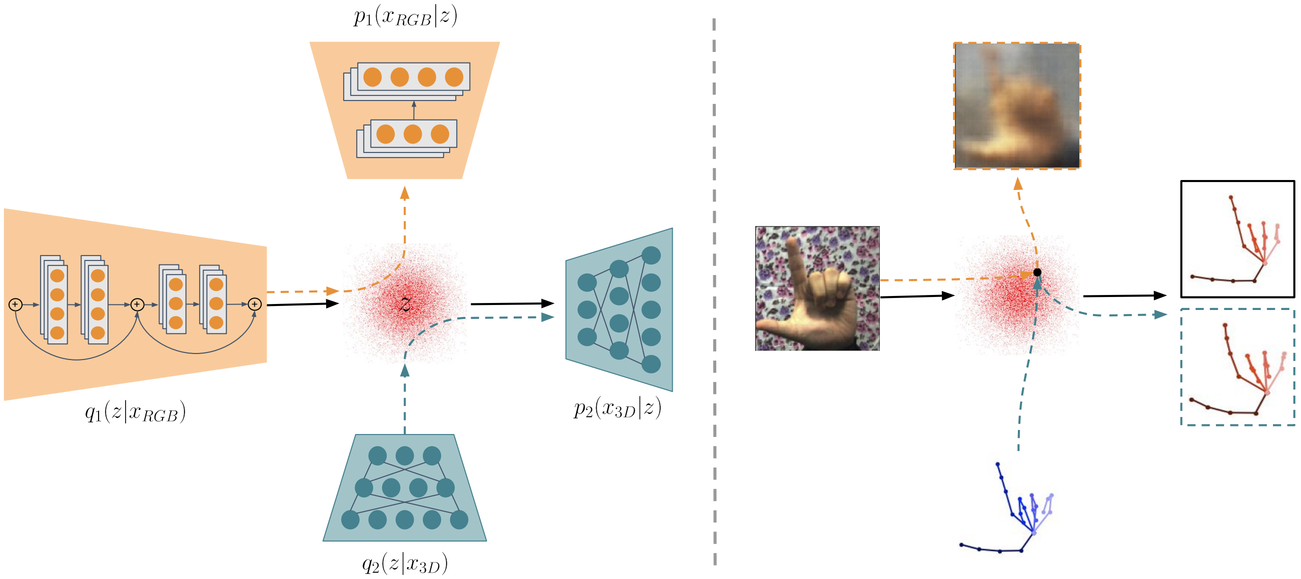


Figure 2: **Schematic overview of our architecture.** Left: a cross-modal latent space z is learned by training pairs of encoder and decoder q, p networks across multiple modalities (e.g., RGB images to 3D hand poses). Auxilliary encoder-decoder pairs help in regularizing the latent space. Right: The approach allows to embed input samples of one set of modalities (here: RGB, 3D) and to produce *consistent* and plausible posterior estimates in several *different* modalities (RGB, 2D and 3D).

imentally show that our methods outperforms [38] both in the 2D-3D lifting setting and the end-to-end hand pose estimation setting, even when using fewer invariances than the original method. Finally, we demonstrate that the same approach can be directly employed to depth images without any modifications to the architecture.

Our work builds on literature in deep generative modeling. Generative Adversarial Nets (GAN) [8] learn an underlying distribution of the data via an adversarial learning process. The Variational Autoencoder (VAE) [11] learns it via optimizing the log-likelihood of the data under a latent space manifold. However unlike GANs, they provide a framework to embed data into this manifold which has been shown to be useful for diverse applications such as multimodal hashing [4]. Aytar et al. [1] use several CNNs to co-embed data from different data modalities for scene classification and Ngiam et al. [14] reconstruct audio and video across modalities via a shared latent space. Our work also aims to create a cross-modal latent space and we provide a derivation of the cross-modal training objective function that naturally admits learning with different data sources all representing physically plausible hand pose configurations.

3. Method

The complex and dexterous articulation of the human hand is difficult to model directly with geometric or physical constraints [18, 30, 33]. However, there is broad agreement in the literature that a large amount of the degrees-of-freedom are not independently controllable and that hand motion, in natural movement, lives in a low-dimensional subspace [20, 31]. Furthermore, it has been shown that dimensionality reduction techniques can provide data-driven priors in RGB-D based hand pose estimation [21, 26]. However, in order to utilize such a low-dimensional sub-space directly for posterior estimation in 3D hand-pose estimation it needs to be i) smooth, ii) continuous and iii) consistent. Due to the inherent difficulties of capturing hand poses, most data sets do not cover the full motion space and hence the desired manifold is not directly attainable via simple dimensionality reduction techniques such as PCA.

We deploy the VAE framework that admits cross-modal training of such a hand pose latent space by using various sources of data representation, even if stemming from different data sets both in terms of input and output. Our cross-modal training scheme, illustrated in Fig. 2, learns to embed hand pose data from different modalities and to reconstruct them either in the same or in a different modality.

More precisely, a set of encoders q take data samples x in the form of either 2D keypoints, RGB or depth images and project them into a low-dimensional latent space z, representing physically plausible poses. A set of decoders p reconstruct the hand configuration in either modality. The focus of our work is on 3D hand pose estimation and therefore on estimating the 3D joint posterior. The proposed approach is fully generative and experimentally we show that it is capable of generating consistent hand configurations across modalities. During training, each input modality alternatively contributes to the construction of the shared latent space. The manifold is continuous and smooth which we show by generating cross-modal samples such as novel pairs of 3D poses and images of natural hands[1].

3.1. Variational Autoencoder

Our cross-modal training objective can be derived from the VAE framework [11], a popular class of generative models, typically used to synthesize data. A latent representation is attained via optimizing the so-called variational lower bound on the log-likelihood of the data:

$$\log p(x) \geq E_{z\sim q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x)||p(z)) \tag{1}$$

Here $D_{KL}(\cdot)$ is the Kullback-Leibler divergence, and the conditional probability distributions $q(z|x)$, $p(x|z)$ are the encoder and decoders, parametrized by neural networks. The distribution $p(z)$ is the prior on the latent space, modeled as $\mathcal{N}(z|0, I)$. The encoder returns the mean μ and variance σ^2 of a normal distribution, such that $z \sim \mathcal{N}(\mu, \sigma^2)$.

In this original form VAEs only take a single data distribution into account. To admit cross-modal training, at least two data modalities need to be considered.

3.2. Cross-modal Hand Pose Latent Space

Our goal is to guide the cross-modal VAE into learning a lower-dimensional latent space of hand poses with the above mentioned desired properties and the ability to project any modality into z and to generate posterior estimates in any modality. For this purpose we re-derive a new objective function for training which leverages multiple modalities. We then detail our training algorithm based on this objective function.

For brevity we use a concrete example in which a data sample x_i (e.g., an RGB image) is embedded into the latent space to obtain the embedding vector z, from which a corresponding data sample x_t is reconstructed (e.g., a 3D joint configuration). To achieve this, we maximize the log-probability of our desired output modality x_t under our model $\log p_\theta(x_t)$, where θ are the model parameters. We will omit the model parameters to reduce clutter.

Similar to the original derivation [11], we start with the quantity $\log p(x_t)$ that we want to maximize:

$$\log p(x_t) = \int_z q(z|x_i) \log p(x_t) dz, \tag{2}$$

exploiting the fact that $\int_z q(z|x_i)dz = 1$ and expanding $p(x_t)$ gives:

$$\int_z q(z|x_i) \log \frac{p(x_t)p(z|x_t)q(z|x_i)}{p(z|x_t)q(z|x_i)} dz. \tag{3}$$

Remembering that $D_{KL}(p(x)||q(x)) = \int_x p(x) \log \frac{p(x)}{q(x)}$ and splitting the integral of Eq (3) we arrive at:

$$\begin{aligned}&\int_z q(z|x_i) \log \frac{q(z|x_i)}{p(z|x_t)} dz + \int_z q(z|x_i) \log \frac{p(x_t)p(z|x_t)}{q(z|x_i)} dz \\ &= D_{KL}(q(z|x_i)||p(z|x_t)) + \int_z q(z|x_i) \log(\frac{p(x_t|z)p(z)}{q(z|x_i)})dz.\end{aligned} \tag{4}$$

Here $p(z|x_t)$ corresponds to the desired but inaccessible posterior, which we approximate with $q(z|x_i)$.

Since $p(x_t)p(z|x_t) = p(x_t|z)p(z)$ and because $D_{KL}(p(x)||q(x)) \geq 0$ for any distribution p, q, we attain the final lower bound:

$$\begin{aligned}&D_{KL}(q(z|x_i)||p(z|x_t)) + \int_z q(z|x_i) \log(\frac{p(x_t|z)p(z)}{q(z|x_i)})dz \\ &\geq \int_z q(z|x_i) \log p(x_t|z) dz - \int_z q(z|x_i) \log \frac{q(z|x_i)}{p(z)} dz \\ &= \mathbb{E}_{z\sim q(z|x_i)}[\log p(x_t|z)] - D_{KL}(q(z|x_i)||p(z)).\end{aligned} \tag{5}$$

Note that we changed signs via the identity $-\log(x) = \log(\frac{1}{x})$. Here $q(z|x_i)$ is our encoder, embedding x_i into the latent space and $p(x_t|z)$ is the decoder, which transforms the latent sample z into the desired representation x_t.

The derivation shows that input samples x_i and target samples x_t can be decoupled via a joint embedding space z where i and t can represent any modality. For example, to maximize $\log p(x_{\text{3D}})$ when given x_{RGB}, we can train with $q(z|x_{\text{RGB}})$ as our encoder and $p(x_{\text{3D}}|z)$ as the decoder.

Importantly the above derivation also allows to train additional encoder-decoder pairs such as $(q(z|x_{\text{RGB}}), p(x_{\text{RGB}}|z))$, at the same time, for the same z. This cross-modal training regime results in a single latent space that allows us to embed and reconstruct multiple data modalities, or even train in a unsupervised fashion.

In the context of hand pose estimation, $p(z)$ represents a hand pose manifold which can be better defined with additional input modalities such as x_{RGB}, x_{2D}, x_{3D}, and even x_{Depth} used in combination.

3.3. Network Architecture

In practice, the encoder q_k for data modality k returns the mean μ and variance σ^2 of a normal distribution for a given sample, from which the embedding z is sampled, i.e $z \sim \mathcal{N}(\mu, \sigma^2)$. However, the decoder p_l directly reconstructs the latent sample z to the desired data modality l.

Fig. 2, illustrates our proposed architecture for the case of RGB based handpose estimation. In this setting we use two encoders for RGB images and 3D keypoints respectively. Furthermore, the architecture contains two decoders for RGB images and 3D joint configurations.

[1] Generated images are legible but blurry. Creating high quality natural images is a research topic in itself.

3.4. Training Procedure

Our cross-modal objective function (Eq 3) follows the training procedure given as pseudo-code in Alg.1. The procedure takes a set of modalities P_{VAE} with corresponding encoders and decoders q_i, p_j, where i, j signify the respective modality, and trains all such pairs iteratively for $\mathcal{E}$ epochs. Note that the embedding space z is always the same and hence we attain a joint cross-modal latent space from this procedure (cf. Fig. 1).

Algorithm 1 Cross-modal Variational Autoencoders

```
$P_{VAE} \leftarrow \{(q_{k_1}, p_{l_1}), (q_{k_2}, p_{l_2}), ...\}$ Encoder/Decoder pairs, where $q_{k_1}$ encodes data from modality $k_1$ and $p_{l_1}$ reconstructs latent samples to data of modality $l_1$.
$\mathcal{E}$ Number of epochs
$e \leftarrow 0$
for $e < \mathcal{E}$ do
    for $(q_k, p_l) \in P_{VAE}$ do
        $x_k, x_l \leftarrow X_k, X_l$ Sample data pair of modality $k, l$
        $\mu, \sigma \leftarrow q_k(x_k)$
        $z \sim \mathcal{N}(\mu, \sigma)$
        $\hat{x}_l \leftarrow p_l(z)$
        $\mathcal{L}_{MSE} \leftarrow ||x_l - \hat{x}_l||_2$
        $\mathcal{L}_{KL} \leftarrow -0.5 * (1 + \log(\sigma^2) - \mu^2 - \sigma^2)$
        $\theta_{q_k} \leftarrow \theta_{q_k} - \nabla_{\theta_{q_k}}(\mathcal{L}_{MSE} + \mathcal{L}_{KL})$
        $\theta_{p_l} \leftarrow \theta_{p_l} - \nabla_{\theta_{p_l}}(\mathcal{L}_{MSE} + \mathcal{L}_{KL})$
    end for
    $e \leftarrow e + 1$
end for
```

4. Experiments

To evaluate the performance of the cross-modal VAE we systematically evaluate the utility of the proposed training algorithm and the resulting cross-modal latent space. This is done via estimation of 3D hand joint positions from three entirely different input modalities: 1) 2D joint locations; 2) RGB image; 3) depth images. In our experiments we explored combinations of different modalities during training. We always predict at least the 3D hand configuration but add further modalities. More specifically we run experiments with the following four **variants**: a) ***Var. 1***: $(x_i \rightarrow x_t)$ b) ***Var. 2***: $(x_i \rightarrow x_t, x_t \rightarrow x_t)$ c) ***Var. 3***: $(x_i \rightarrow x_t, x_i \rightarrow x_i)$ d) ***Var. 4***: $(x_i \rightarrow x_t, x_i \rightarrow x_i, x_t \rightarrow x_t)$, where x_i always signifies the input modality and i takes *one* of the following values: [RGB, 2D, Depth] and t equals the output modality. In our experiments this is always $t = $ 3D but can in general be any target modality. Including the $x_t \rightarrow x_i$ direction neither directly affects the RGB encoder, nor the 3D joint decoder and hence was dropped from our analysis.

4.1. Implementation details

We employ Resnet-18 [9] for the encoding of RGB and depth images. Note that the model size of this encoder is much smaller compared to prior work that directly regresses 3D joint coordinates [15]. The decoders for RGB and depth consist of a series of (TransposedConv, BatchNorm2D and ReLU)-layers. For the case of 2D keypoint and 3D joint encoders and decoders, we use several (Linear, ReLU)-layers. In our experiments we did not observe much increase in accuracy from more complex decoder architectures. We train our architecture with the ADAM optimizer using a learning rate of 10^{-4}. Exact architecture details and hyperparameters can be found in the supplementary materials.

4.2. Datasets

We evaluate our method in the above settings based on several publicly available datasets. For the input modality of **2D keypoints** and **RGB images** only few annotated datasets are available. We test on the datasets of the Stereo Hand Pose Tracking Benchmark [37] (STB) and the Rendered Hand Pose Dataset (RHD) [38]. STB contains 18k images with resolution of 640×480, which are split into a training set with 15k samples and test set with 3k samples. These images are annotated with 3D keypoint locations and the 2D keypoints are recovered via projecting them with the camera intrinsic matrix. The depicted hand poses contain little self-occlusion and variation in global orientation, lighting etc. and are relatively easy to recover.

RHD is a synthetic dataset with rendered hand images, which is composed of 42k training images and 2.7k evaluation images of size 320×320. Similar to STB, both 2D and 3D keypoint locations are annotated. The dataset contains a much richer variety of viewpoints and poses. The 3D human model is set in front of randomly sampled images from Flickr to generate arbitrary backgrounds. This dataset is considerably more challenging due to variable viewpoints and difficult hand poses at different scales. Furthermore, despite being a synthetic dataset the images contain significant amount of noise and blur and are relatively low-res.

For the **depth** data, we evaluate on the ICVL [27], NYU [32], and MSRA [25] datasets. For NYU, we train and test on viewpoint 1 and all 36 available joints, and evaluate on 14 joints as done in [15, 17, 34] while for MSRA, we perform a leave-one-out cross-validation and evaluate the errors for the 9 models trained as done in [15, 25, 34].

4.3. Evaluation metrics

We provide three different metrics to evaluate the performance of our proposed model under various settings: i) The most common metric used in the 3D hand pose estimation literature is the *mean 3D joint error* which measures the average euclidean distance between predicted joints and

	2D→3D RHD	RGB→3D RHD	RGB→3D STB
Var. 1	17.23	**19.73**	8.75
Var. 2	17.82	19.99	8.61
Var. 3	**17.14**	20.04	**8.56**
Var. 4	17.63	20.35	9.57

Table 1: Variant comparison. Mean EPE given in mm. For explanation of variants, see Sec. 4.

ground truth joints. ii) We also report ***Percentage of Correct Keypoints*** **(PCK)** which returns the mean percentage of predicted joints below an euclidean distance of d from the correct joint location. iii) The hardest metric, which reports the ***Percentage of Correct Frames*** **(PCF)** where *all* the predicted joints are within an euclidean distance of d to its respective GT location. We report this only for depth since it is commonly reported in the literature.

4.4. Comparison of variants

We begin with comparing our variants with each other to determine which performs best and experiment on RHD and STB. On both datasets, we test the performance of our model on the task of regressing the 3D joints from RGB directly. Additionally, we predict the 3D joint locations from given 2D joint locations (dimensionality lifting) on RHD.

Table 5 shows our results on the corresponding task and dataset. The errors are given in mean **end-point-error (EPE)** (median EPE is in the supplementary). Var. 3 outperforms the other variants on two tasks; lifting 2D joint locations to 3D on RHD and regressing 3D joint location directly from RGB on STB. On the other hand, Var. 1 is superior in the task of RGB→3D on RHD. However we note that in general, the individual performance differences are minor. This is to be expected, as we conduct all our experiments within individual datasets. Hence even if multiple modalities are present, they capture the same poses and the same inherent information. This indicates that having a shared latent space for generative purposes does not harm the performance and in certain cases can even enhance it. This may be due to the regularizing effect of introducing multiple modalities.

4.5. Comparison to related work

In this section we perform a qualitative analysis of our performance in relation to prior work for both RGB and depth cases. For this, we pick the best variant of the respective task, as determined in the previous section. For the RGB datasets (RHD and STB), we compare against [38]. To the best of our knowledge, it is the only prior work that addresses the same task as we do. In order to compare fairly, we conduct the same data preprocessing. Importantly, in [38] additional information such as **handedness (H)** and **scale of the hand (S)** are provided at test time. Furthermore, the cropped hands are normalized to a roughly uniform size. Finally, they change the task from predicting the global 3D joint coordinates to estimating a palm-relative, **translation invariant (T)** set of joint coordinates by providing ground truth information of the palm center. In our case, the handedness is provided via a boolean flag directly into the model.

However, in order to assess the influence of our learned hand model we incrementally reduce the reliance on invariances which require access to ground-truth information. These results are shown alongside our main algorithm.

2D to 3D. As a baseline experiment we compare our method to that of [38] in the task of lifting 2D keypoints into a 3D hand pose configuration on the RHD dataset. Recently [12] report that given a good 2D keypoint detector, lifting to 3D can yield surprisingly good results, even with simple methods in the case of 3D human pose estimation. Hand pose estimation is considerably more challenging task due to the more complex motion and flexibility of the human hand. Furthermore, [38] provide a separate evaluation of their lifting component which serves as our baseline.

The first column of Table 6 summarizes the mean squared end-point errors (EPE) for the RHD dataset. In general, our proposed model outperforms [38] by a relatively large margin. The bottom rows of Table 6 show results of ours *without* the handedness invariance (H) and the scale invariance (S), we still surpass the accuracy of [38]. This suggests that our model indeed encodes physically plausible hand poses and that reconstructing the posterior from the embedding aids the hand pose estimation task.

RGB to 3D. Here, we evaluate our method on the task of directly predicting 3D hand pose from RGB images, without intermediate 2D keypoint extraction. We run our model and [38] on cropped RGB images for fair comparison.

Zimmermann et al. [38], in which 2D keypoints are first predicted and then lifted into 3D serves as our baseline. We evaluate the proposed model on the STB [37] and RHD [38] datasets. Fig. 10a and 10b show several samples of our prediction on STB and RHD respectively. Even though some images in RHD contain heavily occluded fingers, our method retrieves biomechanically plausible predictions.

The middle column of Table 6 summarizes the results for the harder RHD dataset. Our approachs accuracy exceeds that of [38] by a large margin. Removing available invariances again slightly decreases performance but our models still remains superior to [38]. Looking at the PCK curve comparison in Fig. 4a, we see that our model outperforms [38] for all thresholds.

The rightmost column of Table 6 shows the performance on the STB dataset. The margin of improvement of our approach is considerably smaller. We argue that the perfor-

	2D→3D RHD	RGB→3D RHD	RGB→3D STB
[38] (T+S+H)	22.43	30.42	8.68
Ours (T+S+H)	**17.14**	**19.73**	**8.56**
Ours (T+S)	18.90	20.20	10.16
Ours (T+H)	19.69	22.34	9.59
Ours (T)	21.15	22.53	9.49

Table 2: Related work comparison. Mean EPE given in mm. For explanation of legends, see Sec. 4.5

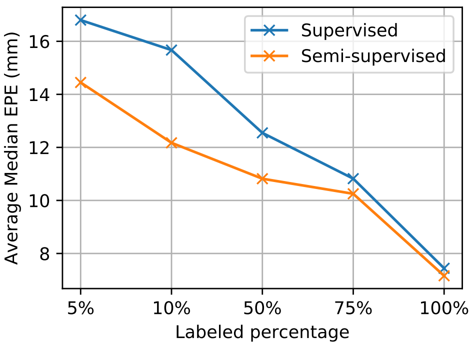
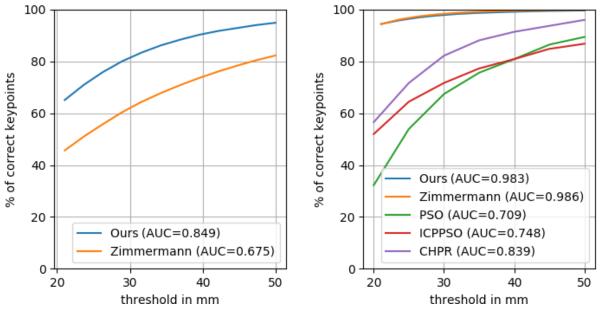
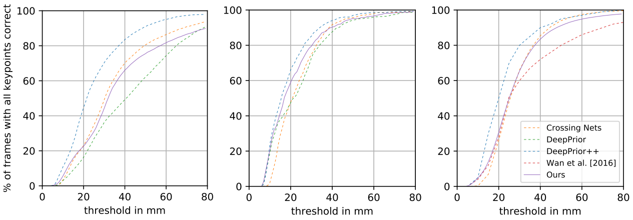


Figure 3: Median EPE of our model trained supervised and semi-supervised as a function of percentage of labeled data.



(a) RHD (b) STB

Figure 4: PCK curve of our best model on RHD and STB for RGB to 3D.



(a) NYU (b) ICVL (c) MSRA

Figure 5: PCF curves for 3D joint estimation from depth input. Our model performs comparably to recent works.

mance on the dataset is saturated as it is much easier (see discussion in Sec. 4.2). Fig. 4b shows the PCK curves on STB, with the other baselines that operate on noisy stereo depth maps and *not* RGB (directly taken from [38]).

Depth to 3D. Given the ready availability of RGB-D cameras, the task of 3D joint position estimation from depth has been explored in great detail and specialized architectures have been proposed. We evaluate our architecture, designed originally for the RGB case, on the ICVL [27], NYU [32] and MSRA [25] datasets. Despite the lower model capacity, our method performs comparably (see Fig. 5) to recent works [15, 17, 34, 35] with just a modification to take 1-channel images as input compared to our RGB case.

4.6. Semi-supervised learning

Due to the nature of cross-training, we can exploit complementary information from additional data. For example, if additional unlabeled images are available, our model can make use of these via cross-training. This is a common scenario, as unlabeled data is plentiful. If not available, acquiring this is by far simpler than recording training data.

To explore this semi-supervised setting, we perform an additional experiment on STB. We simulate a situation where we have labeled and unlabeled data by discarding different percentages of 3D joint data from our dataset. Fig. 3, compares the median EPE of Var. 1 (which can only be trained supervised) with Var. 3 (trained semi-supervised). We see that as more unlabeled data becomes available, Var. 3 can make use of this additional information and improve prediction accuracy up to 22%.

4.7. Generative capabilities

Our model is guided to learn a manifold of hand poses. In this section, we demonstrate the smoothness and consistency of it. To this end, we perform a walk on one dimension of the latent space by embedding two RGB images of separate hand poses into the latent space and obtain two corresponding samples z_1 and z_2. We then decode the latent space samples that reside on the interpolation line between them using our models for RGB and 3D joint decoding. Fig. 6 shows the resulting reconstructions, demonstrating consistency between both decoders. The fingers move in synchrony and the generated synthetic samples are both physically plausible and consistent across modalities. This demonstrates that the learned latent space is indeed smooth and represents a valid statistical model of hand poses.

The smoothness property of the unified latent space is attractive in several regards. Foremost because this potentially enables generation of *labeled* data which in turn may be used to improve current models. Fully exploring this aspect is subject to further research.

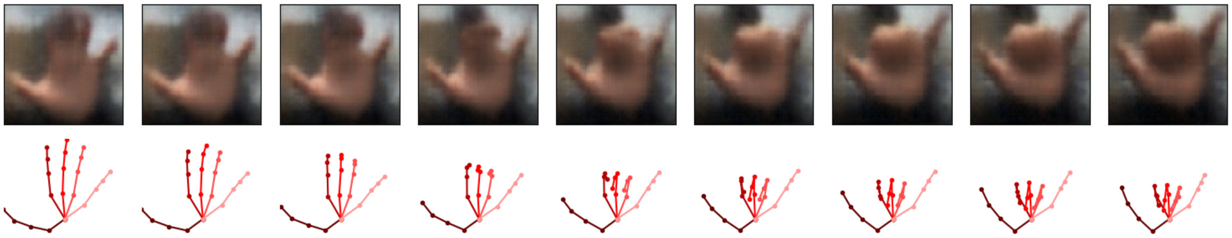
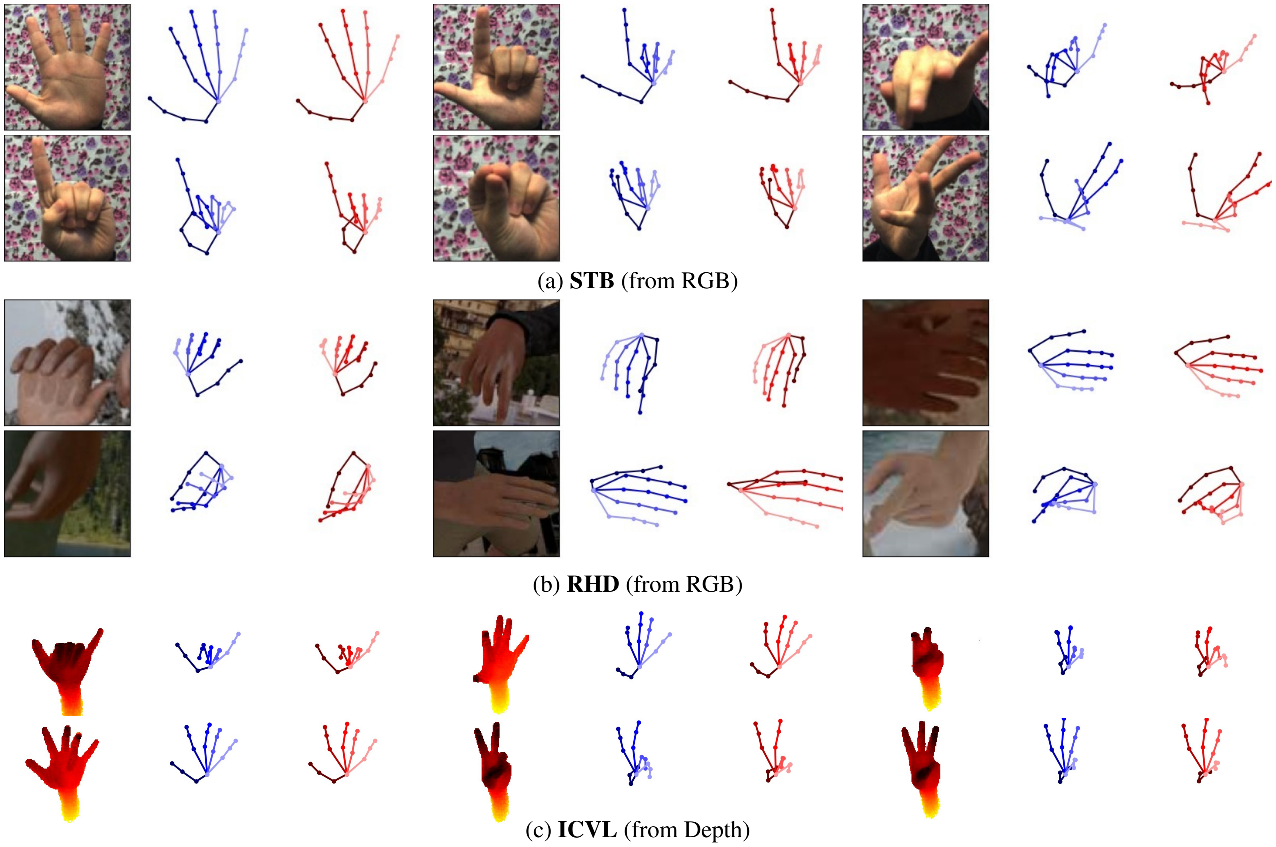


Figure 6: **Latent space walk**. Example of reconstructing samples of the latent space into multiple modalities. The left-most and right-most figures are reconstruction from latent space samples of two real RGB images. The figures in-between are multi-modal reconstruction from interpolated latent space samples, hence are completely synthetic.



(a) **STB** (from RGB)

(b) **RHD** (from RGB)

(c) **ICVL** (from Depth)

Figure 7: **3D joint predictions**. For each triplet, the left most column corresponds to the input image, the middle column is the ground truth 3D joint skeleton and the right column is our corresponding prediction.

5. Conclusion

We have proposed a new approach to estimate 3D hand pose configurations from RGB and depth images. Our approach is based on a re-derivation of the variational lower bound that admits training of several independent pairs of encoders and decoders, shaping a joint cross-modal latent space representation. We have experimentally shown that the proposed approach outperforms the state-of-the art on publicly available RGB datasets and is at least comparable to highly specialized state-of-the-art methods on depth data. Finally, we have shown the generative nature of the approach which suggests that we indeed learn a usable and physically plausible statistical hand model, enabling direct estimation of the 3D joint posterior.

6. Acknowledgements

This work was supported in parts by the ERC grant OPTINT (StG-2016-717054)

References

[1] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba. Cross-modal scene networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3

[2] D. Bouchacourt, P. K. Mudigonda, and S. Nowozin. Disco nets: Dissimilarity coefficients networks. In *Advances in Neural Information Processing Systems*, pages 352–360, 2016. 2

[3] C. Choi, A. Sinha, J. Hee Choi, S. Jang, and K. Ramani. A collaborative filtering approach to real-time hand pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2336–2344, 2015. 2

[4] V. Erin Liong, J. Lu, Y.-P. Tan, and J. Zhou. Cross-modal deep variational hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4077–4085, 2017. 3

[5] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1):52–73, 2007. 2

[6] L. Ge, H. Liang, J. Yuan, and D. Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3593–3601, 2016. 2

[7] L. Ge, H. Liang, J. Yuan, and D. Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1991–2000, 2017. 2

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[10] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun. Real time hand pose estimation using depth sensors. In *Consumer depth cameras for computer vision*, pages 119–137. Springer, 2013. 2

[11] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3, 4

[12] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *International Conference on Computer Vision*, 2017. 6

[13] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. *arXiv preprint arXiv:1704.02201*, 2017. 2

[14] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011. 3

[15] M. Oberweger and V. Lepetit. Deepprior++: Improving fast and accurate 3d hand pose estimation. In *International Conference on Computer Vision Workshops*, 2017. 1, 2, 5, 7, 12

[16] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015. 1, 2

[17] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015. 5, 7

[18] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BmVC*, volume 1, page 3, 2011. 2, 3

[19] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1106–1113, 2014. 2

[20] M. Santello, M. Flanders, and J. F. Soechting. Postural hand synergies for tool use. *Journal of Neuroscience*, 18(23):10105–10115, 1998. 2, 3

[21] M. Schröder, J. Maycock, H. Ritter, and M. Botsch. Real-time hand tracking using synergistic inverse kinematics. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 5447–5454. IEEE, 2014. 3

[22] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3633–3642. ACM, 2015. 2

[23] A. Sinha, C. Choi, and K. Ramani. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4150–4158, 2016. 2

[24] S. Sridhar, A. Oulasvirta, and C. Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2456–2463, 2013. 2

[25] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 824–832, 2015. 2, 5, 7

[26] A. Tagliasacchi, M. Schroeder, A. Tkach, S. Bouaziz, M. Botsch, and M. Pauly. Robust articulated-icp for real-time hand tracking. *Computer Graphics Forum (Proc. Symposium on Geometry Processing)*, 2015. 2, 3

[27] D. Tang, H. Jin Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 2, 5, 7

[28] D. Tang, J. Taylor, P. Kohli, C. Keskin, T.-K. Kim, and J. Shotton. Opening the black box: Hierarchical sampling optimization for estimating human hand pose. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3325–3333, 2015. 2

[29] D. Tang, T.-H. Yu, and T.-K. Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *Proceedings of the IEEE international conference on computer vision*, pages 3224–3231, 2013. 2

[30] J. Taylor, L. Bordeaux, T. Cashman, B. Corish, C. Keskin, T. Sharp, E. Soto, D. Sweeney, J. Valentin, B. Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (TOG)*, 35(4):143, 2016. 2, 3

[31] E. Todorov and Z. Ghahramani. Analysis of the synergies underlying complex hand manipulation. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 2, pages 4637–4640. IEEE, 2004. 3

[32] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):169, 2014. 2, 5, 7

[33] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193, 2016. 2, 3

[34] C. Wan, T. Probst, L. Van Gool, and A. Yao. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 5, 7

[35] C. Wan, A. Yao, and L. Van Gool. Hand pose estimation from local surface normals. In *European Conference on Computer Vision*, pages 554–569. Springer, 2016. 1, 2, 7

[36] Q. Ye, S. Yuan, and T.-K. Kim. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In *European Conference on Computer Vision*, pages 346–361. Springer, 2016. 2

[37] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016. 2, 5, 6

[38] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In *International Conference on Computer Vision*, 2017. 1, 2, 3, 5, 6, 7, 11, 12

Encoder	Decoder		
ResNet-18	Linear(4096)	BatchNorm	ReLU
	Reshape(256, 4, 4)		
	ConvT(128)	BatchNorm	ReLU
	ConvT(64)	BatchNorm	ReLU
	ConvT(32)	BatchNorm	ReLU
	ConvT(16)	BatchNorm	ReLU
	ConvT(8)	BatchNorm	ReLU
	ConvT(3)		

Table 4: Encoder and Decoder architecture for RGB data. ConvT corresponds to a layer performing transposed Convolution. The number indicated in the bracket is the number of output filters. Each ConvT layer uses a 4×4 kernel, stride of size 2 and padding of size 1.

Encoder/Decoder	
Linear(512)	ReLU
Linear(512)	ReLU
Linear(512)	ReLU
Linear(512)	ReLU
Linear(512)	ReLU
Linear(512)	

Table 3: Encoder and decoder architecture.

7. Supplementary

This documents provides additional information regarding our main paper and discusses architecture, training and further implementation details. Furthermore, we provide additional experimental results in particular those that illustrate the benefit of the cross-modal latent space representation.

7.1. Training details

All code was implemented in PyTorch. For all models, we used the ADAM optimizer with its default parameters to train and set the learning rate of 10^{-4}. The batch size was set to 64.

2D to 3D. For the 2D to 3D modality we use identical encoder and decoder architectures, consisting of a series of (Linear,ReLU)-layers. The exact architecture is summarized in table 3.

RGB to 3D. For the RGB to 3D modality, images were normalized to the range $[-0.5, 0.5]$ and we used data augmentation to increase the dataset size. More specifically, we randomly shifted the bounding box around the hand image, rotated the cropped images in the range $[-45°, 45°]$ and applied random flips along the y-axis. The resulting image was then resized to 256×256. The joint data was augmented accordingly.

Because the RHD and STB datasets have non-identical hand joint layouts (RHD gives the wrist-joint location, whereas STB gives the palm-joint location), we shifted the wrist joint of RHD into the palm via interpolating between the wrist and first middle-finger joint. We trained on both hands of the RHD dataset, whereas we used both views of the stereo camera of the STB dataset. This is the same procedure as in [38]. The encoder and decoder architectures for RGB data are detailed in table 4. We used the same encoder/decoder architecture for the 3D to 3D joint modality as for the 2D to 2D case (shown in table 3).

Depth to 3D. We used the same architecture and training regime as for the RGB case. The only difference was adjusting the number of input channels from 3 to 1.

7.2. Qualitative Results

In this section we provide additional qualitative results, all were produced with the architecture and training regime detailed in the main paper.

Latent space consistency. In Fig. 8 we embed data samples from RHD and STB into the latent space and perform a t-SNE embedding. Each data modality is color coded (blue: RGB images, green: 3D joints, yellow: 2D joints). Here, Fig. 8a displays the embedding for our model when it is cross-trained. We see that each data modality is evenly distributed, forming a single, dense, approximately Gaussian cluster. Compared to Fig. 8b which shows the embedding for the same model without cross-training, it is clear that each data modality lies on a separate manifold. This figure indicates that cross-training is vital for learning a multimodal latent space.

To further evaluate this property, in Fig. 9 we show samples from the manifold, decoding them into different modalities. The latent samples are chosen such that the lie on an interpolated line between two embedded images. In other words, we took sample x^1_{RGB} and x^2_{RGB} and encoded them to obtain latent sample z^1 and z^2. We then interpolated linearly between these two latent samples, obtaining latent samples z^j which were then decoded into the 2D, 3D and RGB modality, resulting in a triplet. Hence the left-most and right-most samples of the figure correspond to reconstruction of the RGB image and prediction of its 2D and 3D keypoints, whereas the middle figures are completely synthetic. It's important to note here that each decoded triplet originates from the same point in the latent space. This visualization shows that our learned manifold is indeed consistent amongst all three modalities. This result is in-line with the visualization of the joint embedding space visualized in Fig. 8.

Additional figures. Fig. 10a visualizes predictions on STB. The poses contained in the dataset are simpler, hence the predictions are very accurate. Sometimes the estimated hand poses even appear to be more correct than the ground

	2D→3D RHD	RGB→3D RHD	RGB→3D STB
[38] (T+S+H)	18.84	24.49	7.52
Ours (T+S+H)	**14.46**	**16.74**	**7.16**
Ours (T+S)	14.91	16.93	9.11
Ours (T+H)	16.41	18.99	8.33
Ours (T)	16.92	19.10	7.78

Table 6: The median end-point-error (EPE). Comparison to related work

	2D→3D RHD	RGB→3D RHD	RGB→3D STB
Variant 1	14.68	**16.74**	7.44
Variant 2	15.13	16.97	7.39
Variant 3	**14.46**	16.96	**7.16**
Variant 4	14.83	17.30	8.16

Table 5: The median end-point-error (EPE). Comparing our variants.

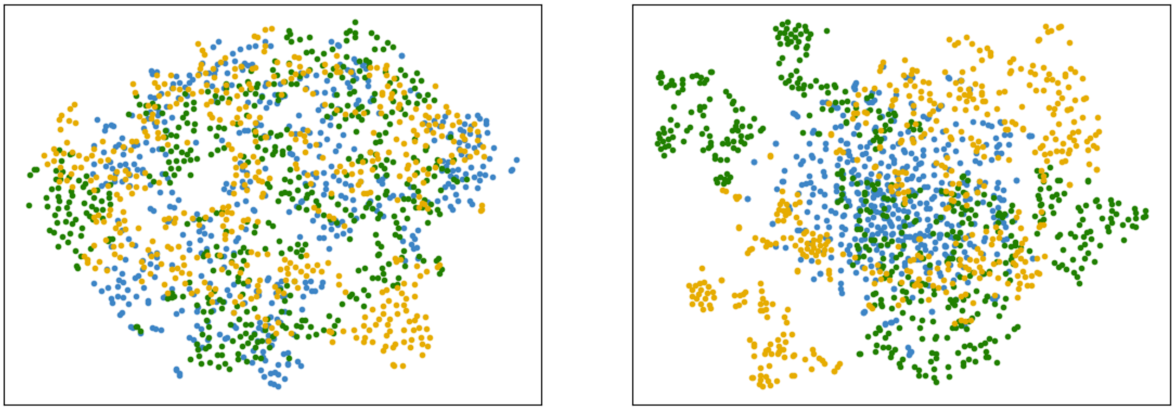
truth (cf. right most column). Fig. 10b shows predictions on RHD. The poses are considerably harder than in the STB dataset and contain more self-occlusion. Nevertheless, our model is capable of predicting realistic poses, even for occluded joints. Fig. 12 shows similar results for depth images.

Fig. 11 displays the input image, its ground truth joint skeleton and predictions of our model. These were constructed by sampling repeatedly from the latent space from the predicted mean and variance which are produced by the RGB encoder. Generally, there are only minor variations in the pose, showing the high confidence of predictions of our model.

7.3. Influence of model capacity

All of our models predicting 3D joint skeleton from RGB images have strictly less parameters than [38]. Our smallest model consists of $12'398'387$ parameters, and the biggest ranges up to $14'347'346$. In comparison, [38] uses $21'394'529$ parameters. Yet, we still outperform them on RHD and reach parity on the saturated STB dataset. This provides further evidence of the proposed approach to learn a manifold of physically plausible hand configurations and to leverage this for the prediction of joint positions directly from an RGB image.

[15] employ a ResNet-50 architecture to predict the 3D joint coordinates directly from depth. In the experiment reported in the main paper, our architecture produced a slightly higher mean EPE (8.5) in comparison to DeepPrior++ (8.1). We believe this can be mostly attributed to differences in model capacity. To show this, we re-ran our experiment on depth images, using the ResNet-50 architecture as encoder and achieved a mean EPE of 8.0.



(a) Cross-trained.

(b) Not cross-trained.

Figure 8: **t-SNE embedding of multi-modal latent space**. The two figures show the embedding of data samples from different modalities (blue: RGB images, green: 3D joints, yellow: 2D joints). In the left figure, our model was cross-trained, whereas in the right figure, each data modality was trained separately. This shows that in order to learn a multi-modal latent space, cross-training is vital.

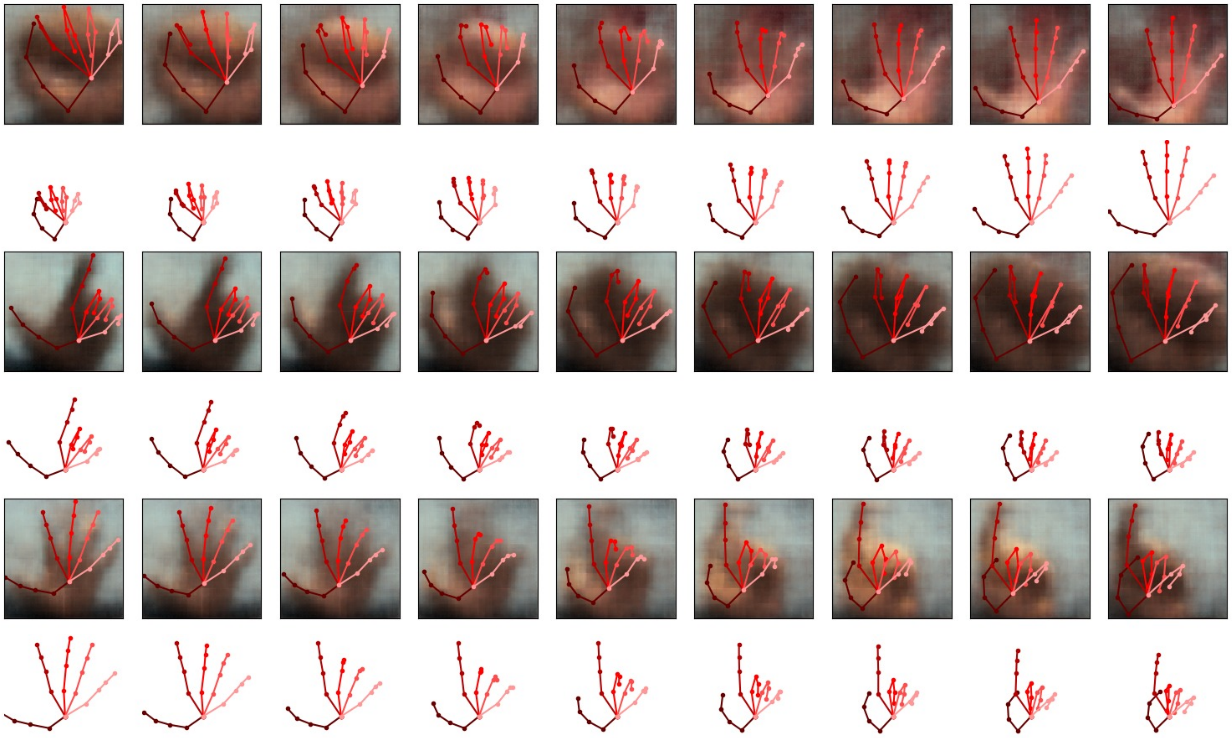
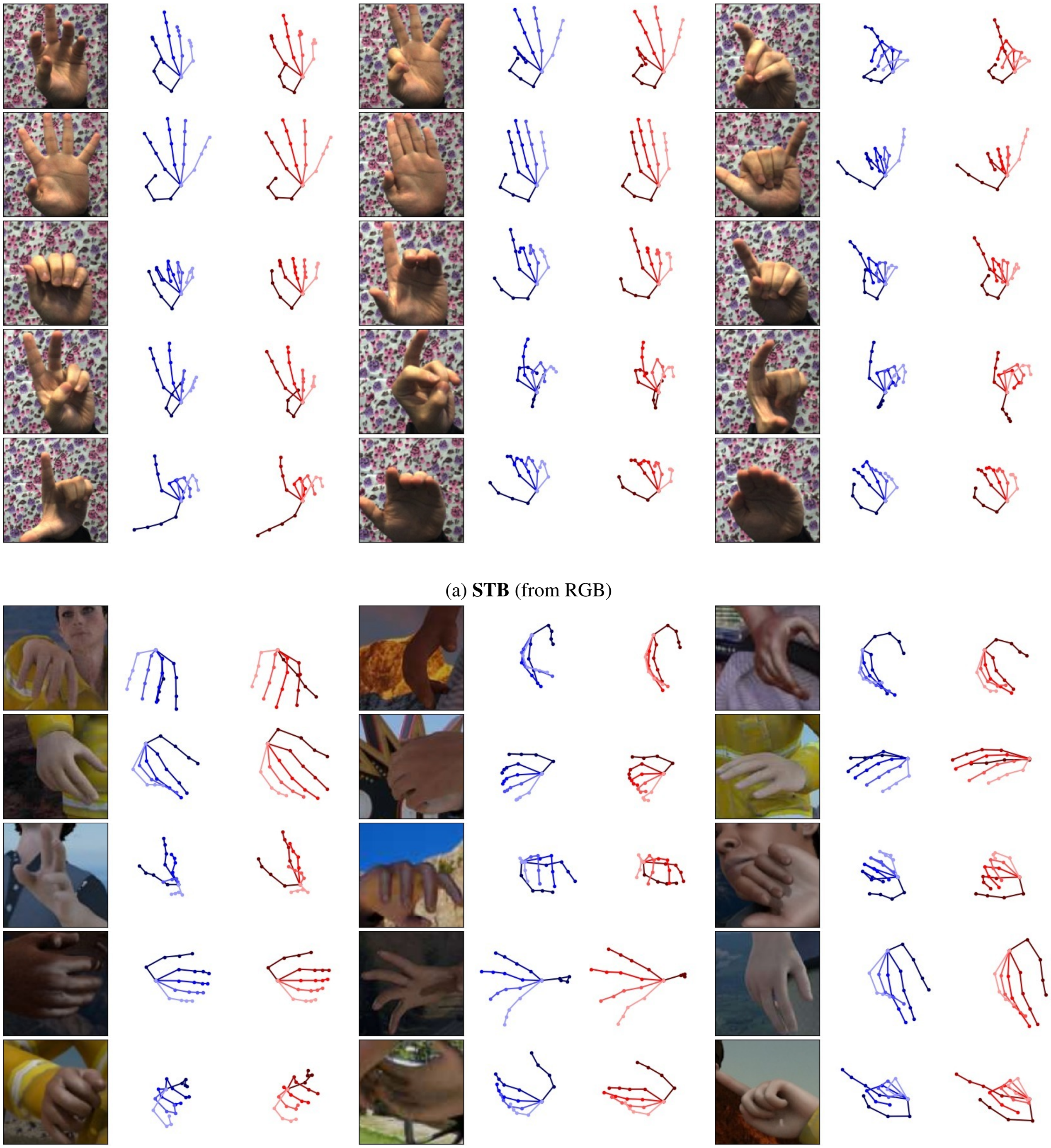


Figure 9: **Latent space walk**. The left-most and right-most figures are reconstruction from latent space samples of two real RGB images. The figures in-between are multi-modal reconstruction from interpolated latent space samples, hence are completely synthetic. Shown are the reconstructed RGB images, with the reconstructed 2D keypoints (overlayed on the RGB image) and the corresponding reconstructed 3D joint skeleton. Each column-triplet is created from the same point in the latent space.



(a) **STB** (from RGB)

(b) **RHD** (from RGB)

Figure 10: **RGB to 3D joint prediction**. Blue is ground truth and red is the prediction of our model.

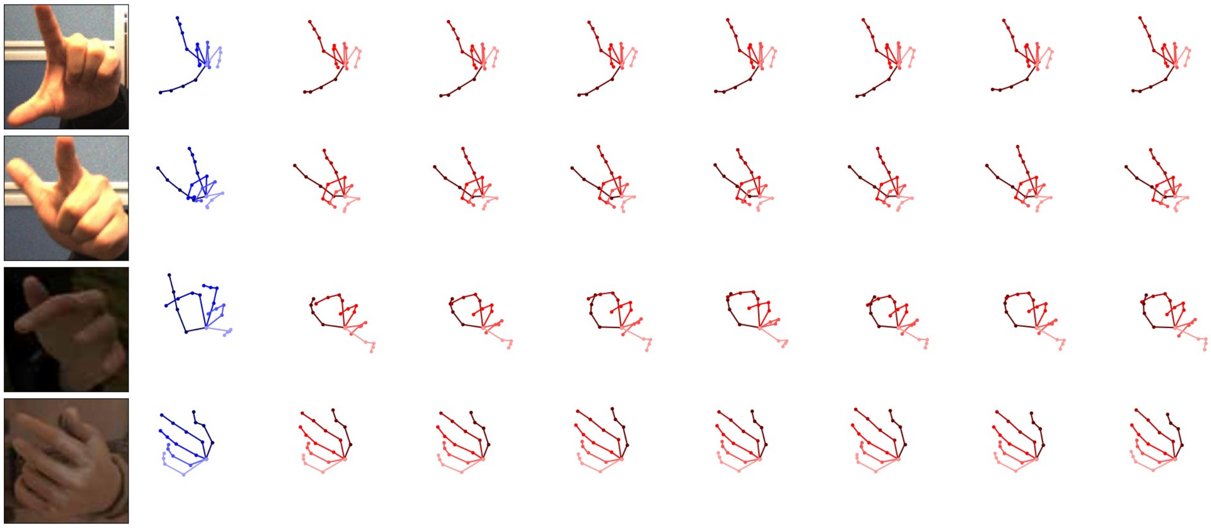


Figure 11: **Sampling from prediction**. This figure shows the resulting reconstruction from samples $z \sim \mathcal{N}(\mu, \sigma^2)$ (red), where μ, σ^2 are the predicted mean and variance output by the RGB encoder. Ground-truth is provided in blue for comparison.

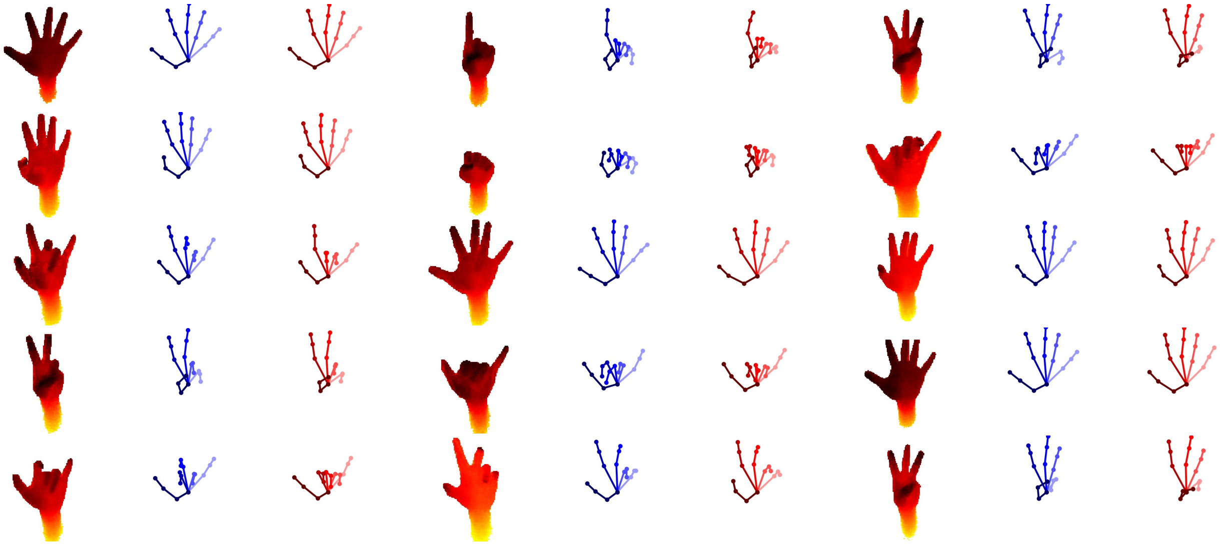


Figure 12: **Depth to 3D joint predictions**. For each row-triplet, the left most column corresponds to the input image, the middle column is the ground truth 3D joint skeleton and the right column is our corresponding prediction.