# Supplementary:
# Cross-modal Deep Variational Hand Pose Estimation

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

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

Table 1: Encoder and decoder architecture.

| 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 2: 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 $4 \times 4$ kernel, stride of size 2 and padding of size 1.

## 1. 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.

### 1.1. 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 1.

**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 [2]. The encoder and decoder architectures for RGB data are detailed in table 2. We used the same encoder/decoder architecture for the 3D to 3D joint modality as for the 2D to 2D case (shown in table 1).

**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.

### 1.2. 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. 1 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. 1a 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. 1b 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 multi-modal latent space.

To further evaluate this property, in Fig. 2 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. 1.

**Additional figures.** Fig. 3a 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 truth (cf. right most column). Fig. 3b 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. 5 shows similar results for depth images.

Fig. 4 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.

### 1.3. Influence of model capacity

All of our models predicting 3D joint skeleton from RGB images have strictly less parameters than [2]. Our smallest model consists of $12'398'387$ parameters, and the biggest ranges up to $14'347'346$. In comparison, [2] 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.

[1] 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.

| | 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 3: The median end-point-error (EPE). Comparing our variants.

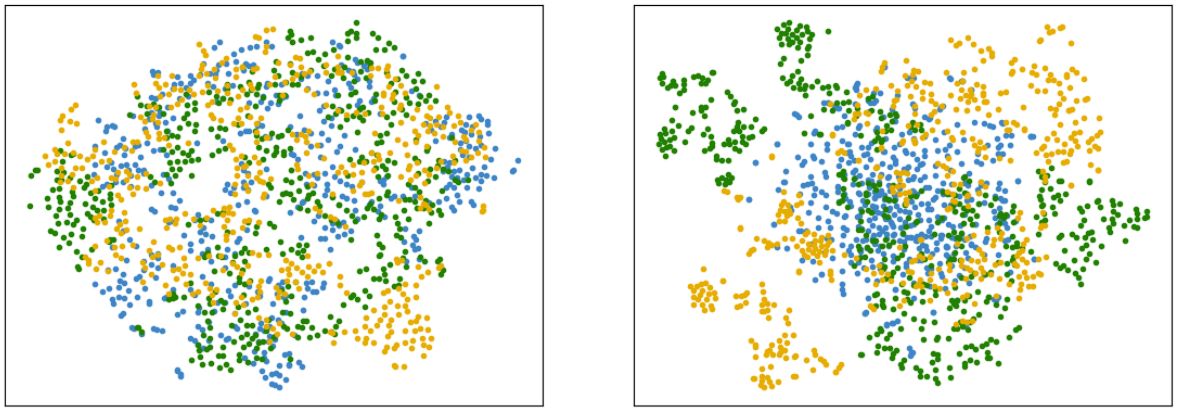
| | 2D→3D RHD | RGB→3D RHD | RGB→3D STB |
|---|---|---|---|
| [2] (T+S+H) | 18.84 | 24.49 | 7.52 |
| Ours (T+S+H) | **14.46** | **16.74** | **7.16** |
| Ours (T+S) | 14.91 | 16.93 | 9.11 |
| Ours (T+H) | 16.41 | 18.99 | 8.33 |
| Ours (T) | 16.92 | 19.10 | 7.78 |

Table 4: The median end-point-error (EPE). Comparison to related work

## References


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

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



(a) Cross-trained. (b) Not cross-trained.

Figure 1: **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.

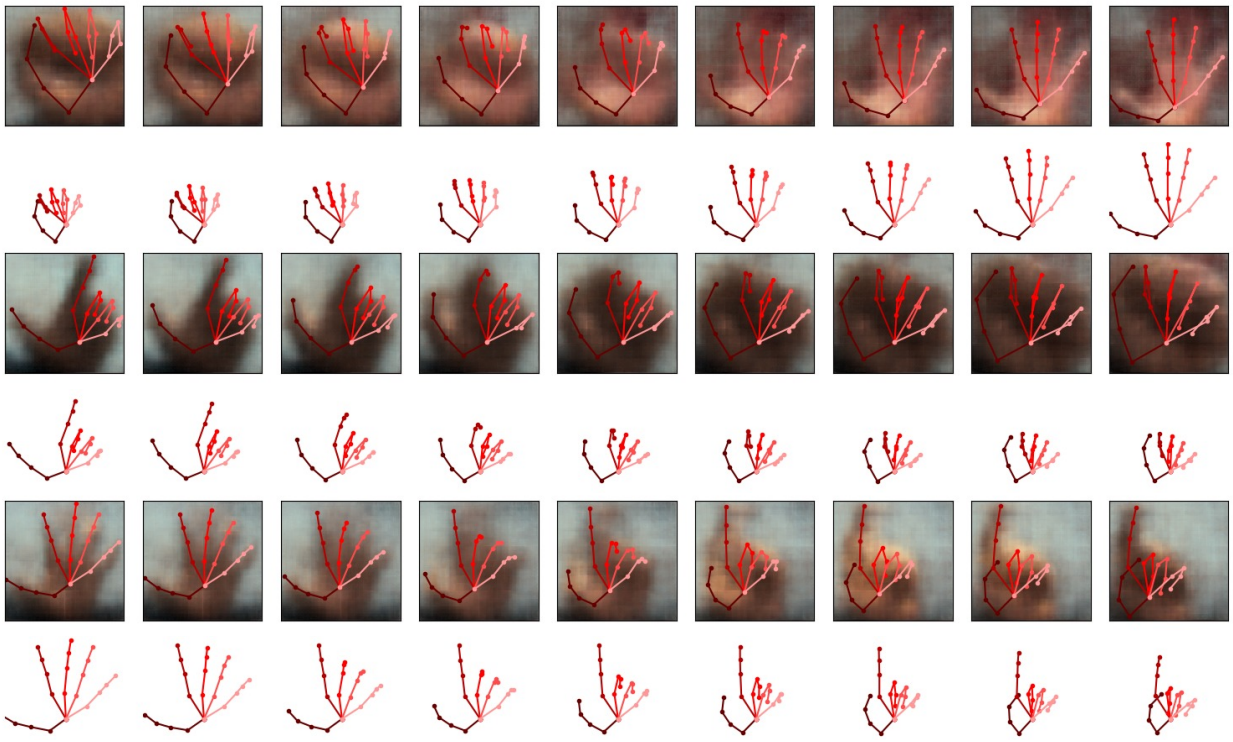
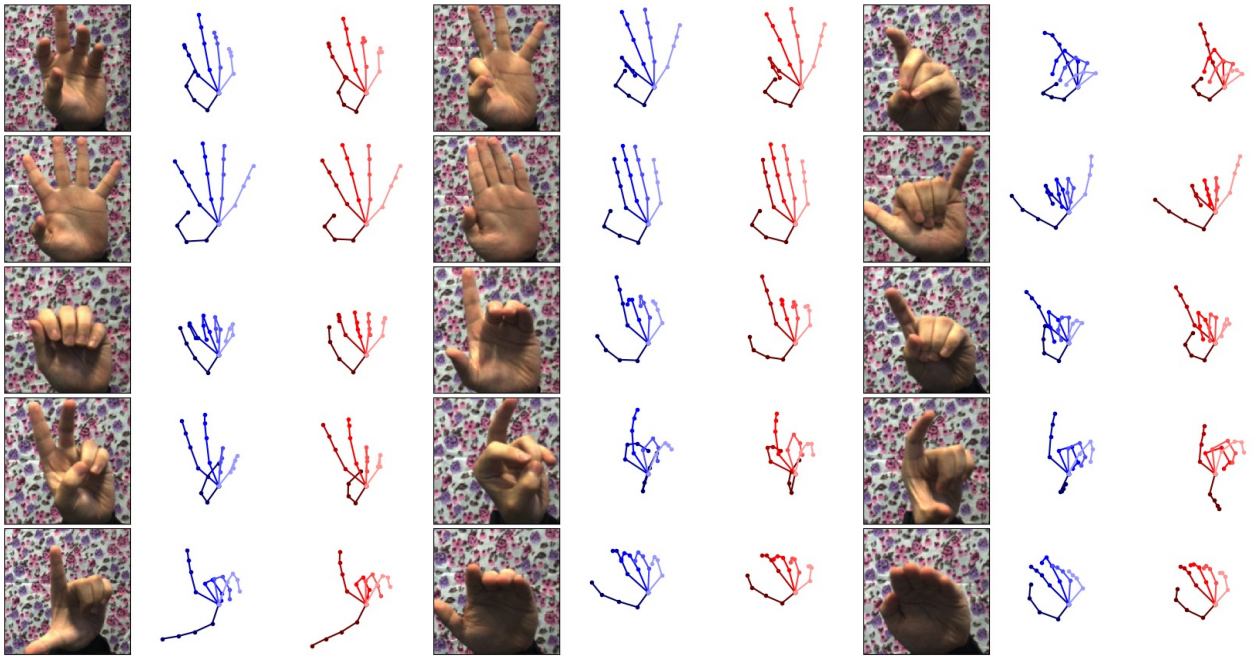
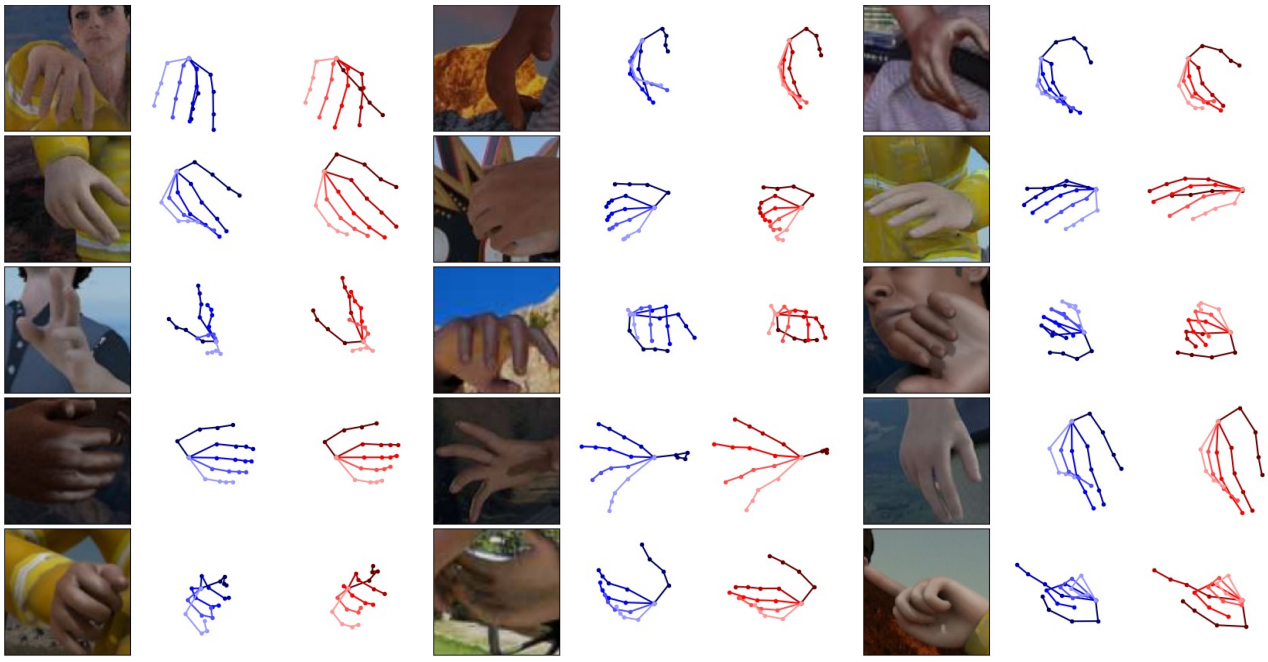


Figure 2: **Latent space walk**. The left-most and right-most figures are reconstruction from latent space samples of two real RGB images. The figures in-between are multi-modal reconstruction from interpolated latent space samples, hence are completely synthetic. Shown are the reconstructed RGB images, 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 3: **RGB to 3D joint prediction**. Blue is ground truth and red is the prediction of our model.

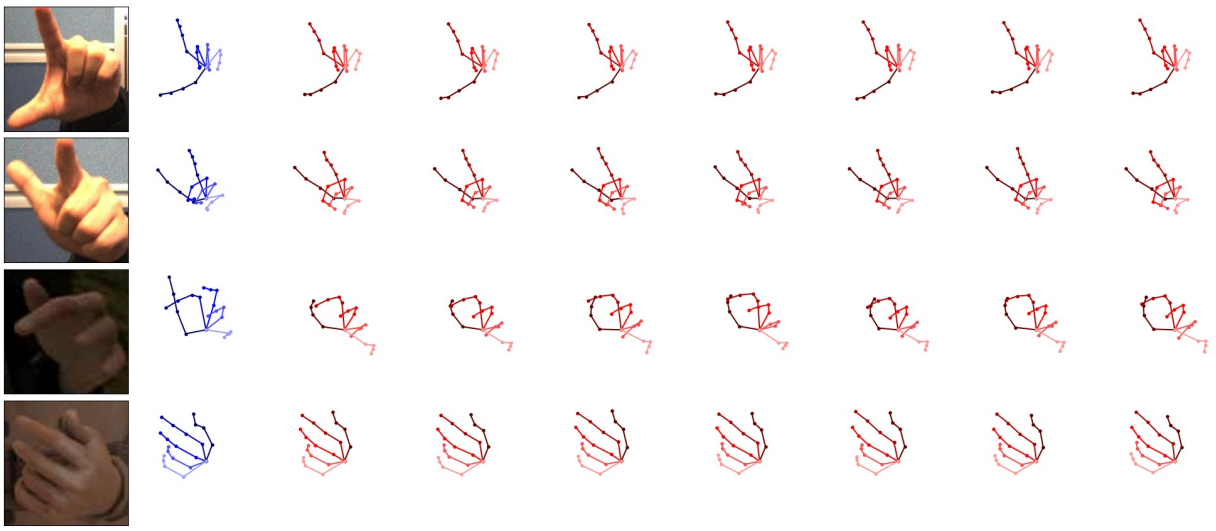


Figure 4: **Sampling from prediction**. This figure shows the resulting reconstruction from samples $z \sim \mathcal{N}(\mu, \sigma^2)$ (red), where $\mu, \sigma^2$ are the predicted mean and variance output by the RGB encoder. Ground-truth is provided in blue for comparison.

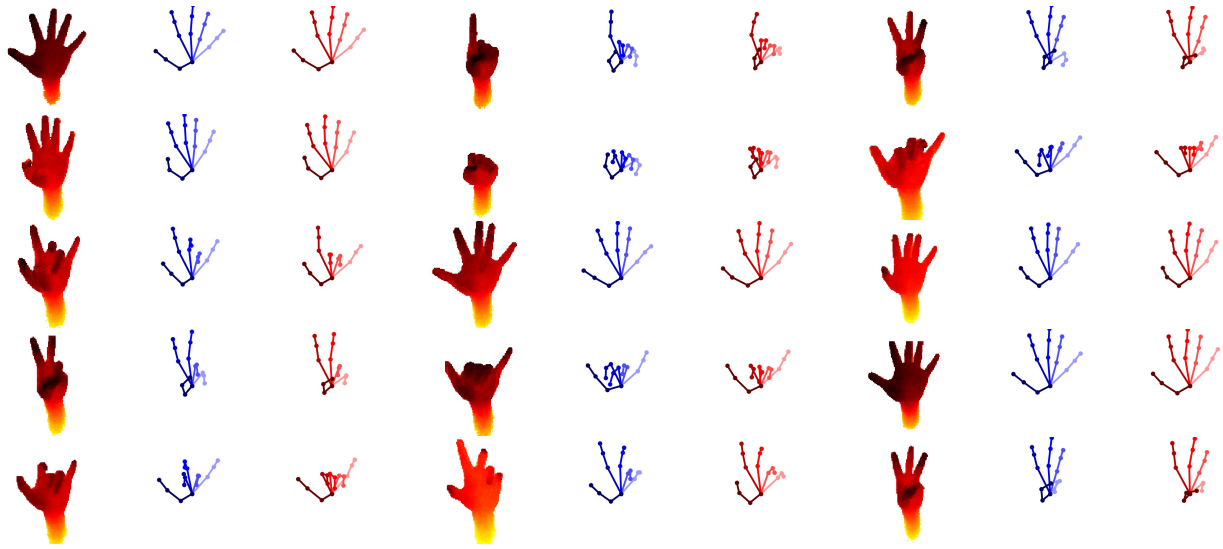


Figure 5: **Depth to 3D joint predictions**. For each row-triplet, the left most column corresponds to the input image, the middle column is the ground truth 3D joint skeleton and the right column is our corresponding prediction.