Conditional Hand Image Generation using Latent Space Supervision in Random Variable Variational Autoencoders

Vassilis C. Nicodemou[1,2] Iason Oikonomidis[2] Giorgos Karvounas[1,2] Antonis Argyros[1,2]

[1]Computer Science Department, University of Crete, Heraklion, Greece
[2]Institute of Computer Science, FORTH, Heraklion, Greece

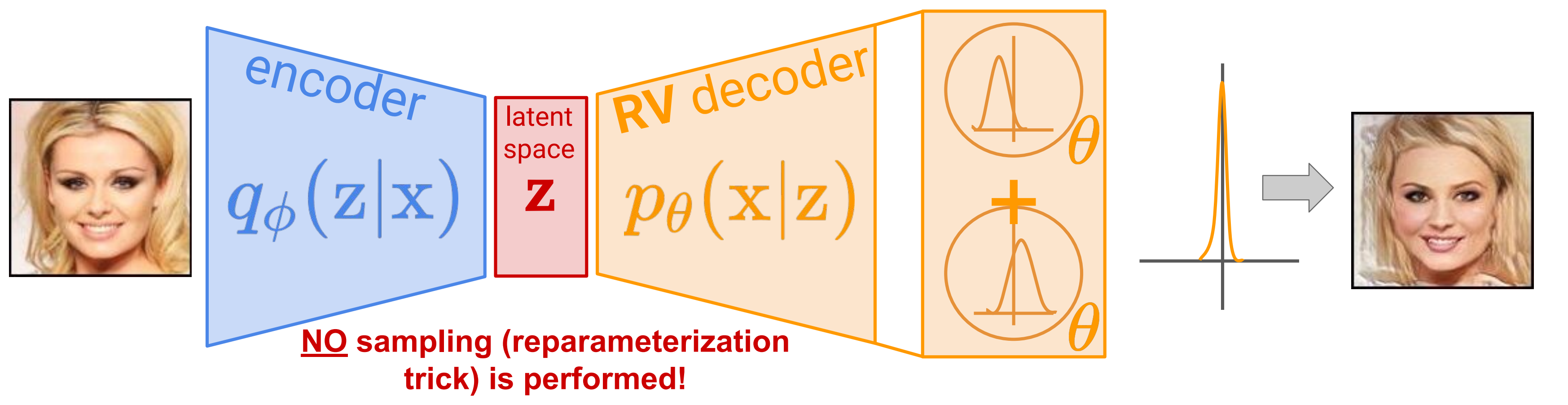
Summary

We introduce a new **generative method** that generates **RGB images** of **hands** given **specific poses** and **arbitrary appearances**. We achieve this by utilizing the strength of **RV-VAEs** and **supervising** a **part of** their **latent space**, while leaving the **rest unsupervised** during training. This results in the creation of **realistic hand images** even on **unseen poses** during inference that can be beneficial in **enhancing hand datasets** or even **extracting appearance attributes**.

Motivation

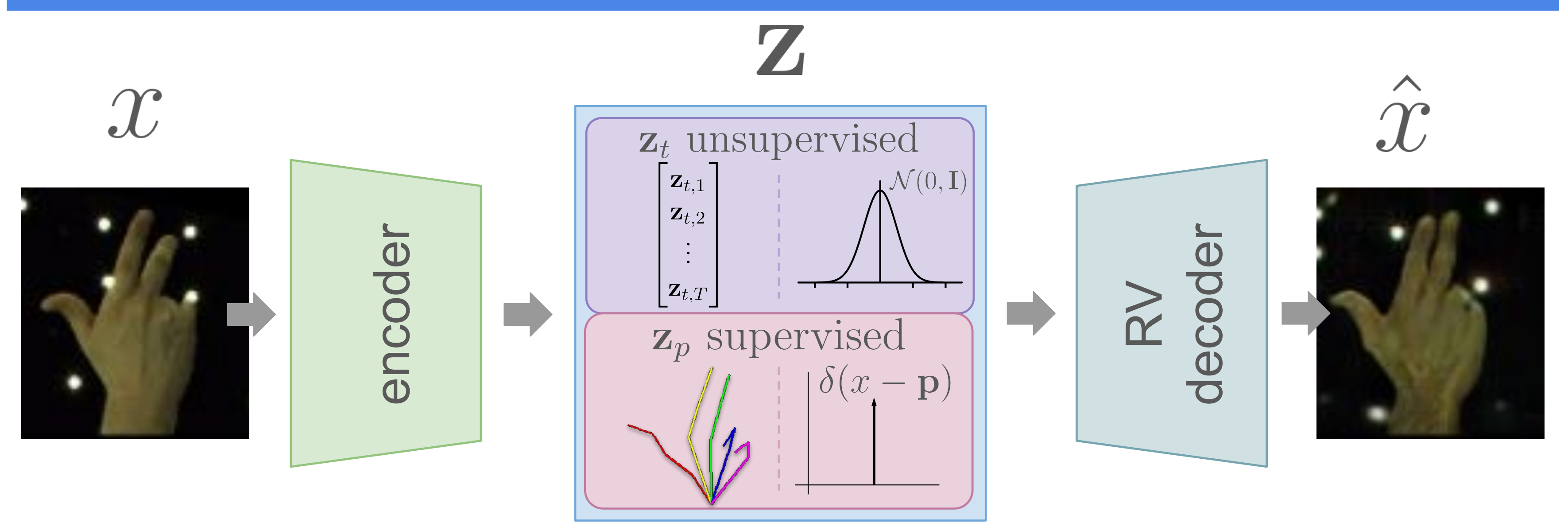
Control over **pose** and **appearance** of generated images. **Map** those two on **specific spaces** that are **easy** to **traverse/navigate** and are **accurately structured**
→ **RV-VAE latent space**

Random Variable VAE


encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
latent space
$\mathbf{z}$
RV decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
θ
θ
NO sampling (reparameterization trick) is performed!

No sampling → more **accurate representation of latent space**. We use the **whole distributions** that are the output of the encoder (in our case one distribution for **pose** and one for **appearance**).

Supervised RV-VAE



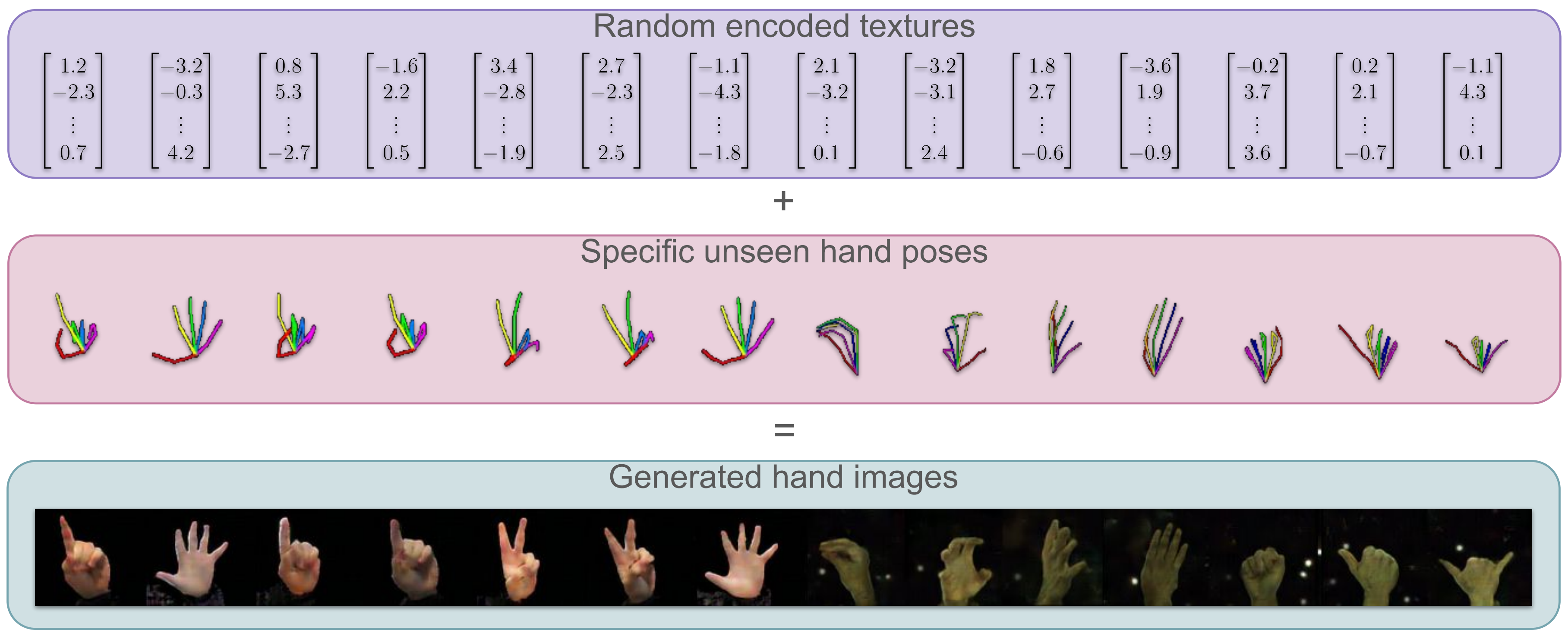
$$\underbrace{\mathcal{L}(\theta,\phi;\mathbf{x})}_{\text{ELBO Loss}} = \underbrace{-D_{KL}(q_\phi(\mathbf{z}_t|\mathbf{x})||p(\mathbf{z}_t))}_{\text{Unsupervised KL divergence for appearance (texture)}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x}),q_\phi(\mathbf{z}_p|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{Reconstruction error MSE w.r.t. pose and texture}} + \underbrace{\mathbb{E}_{p(\mathbf{z}_p|\mathbf{x})}[\log q_\phi(\mathbf{z}_p|\mathbf{x})]}_{\text{Supervised MSE between the estimated and g.t. pose}}$$

Encoder outputs **two random variables**:
➢ The **unsupervised** $\mathbf{z}_t \sim \mathcal{N}(0,\mathbf{I})$
➢ The **supervised** $\mathbf{z}_p \sim \delta(\mathbf{p})$
During **training** they are **sent** both to the **RV decoder** to **synthesize** the input **image** x

For more information, contact: {nikodim, oikonom, gkarv, argyros}@ics.forth.gr

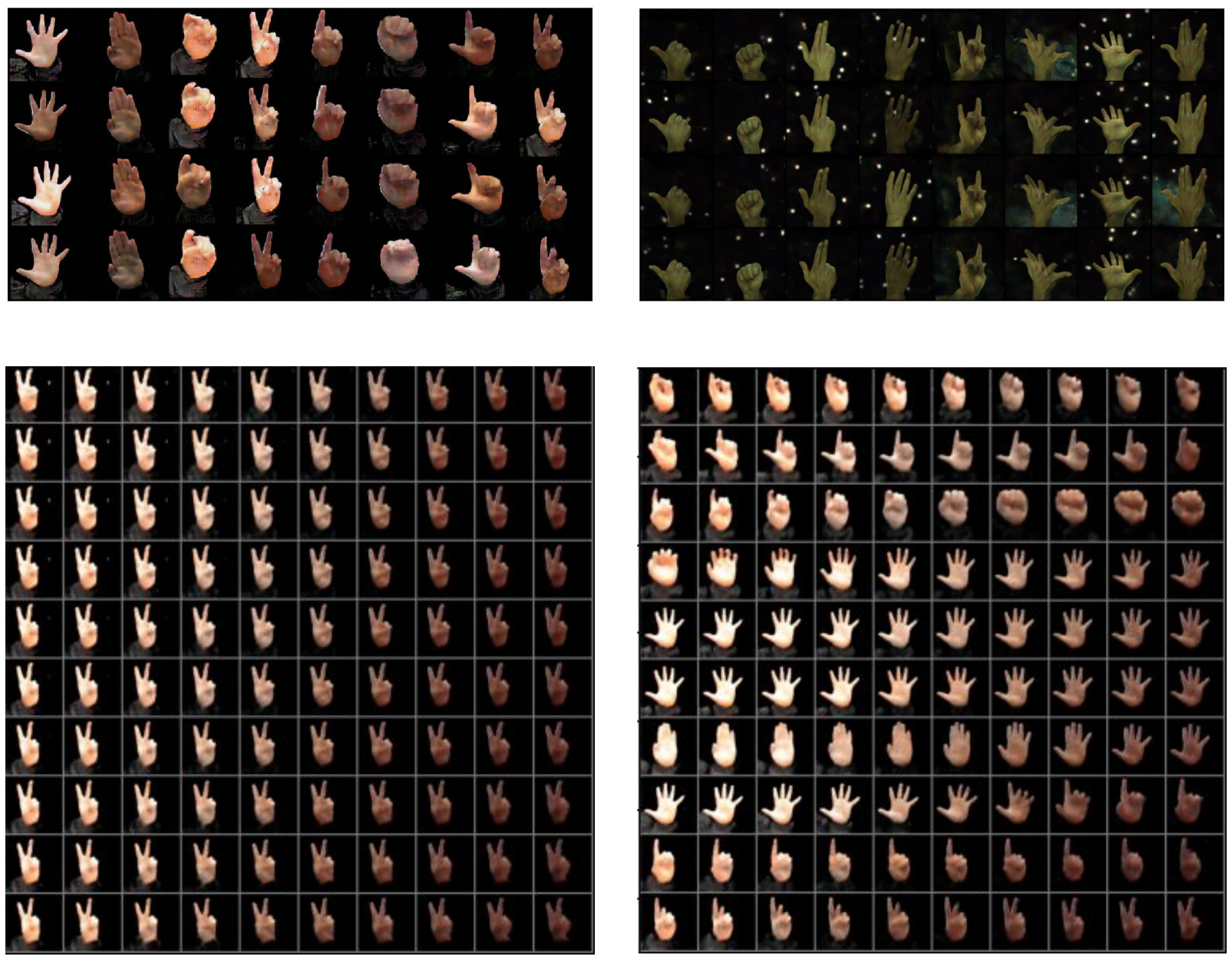
Inference in SRV-VAE

For **inference** we need a **random appearance vector** and a **specific pose** to synthesize **new hand images**.



Experimental Results

Qualitative results by **fixing** the **pose** and **changing** the **appearance** randomly or linearly.



FID values for generated images of:
➢**Test poses** with **test appearances**

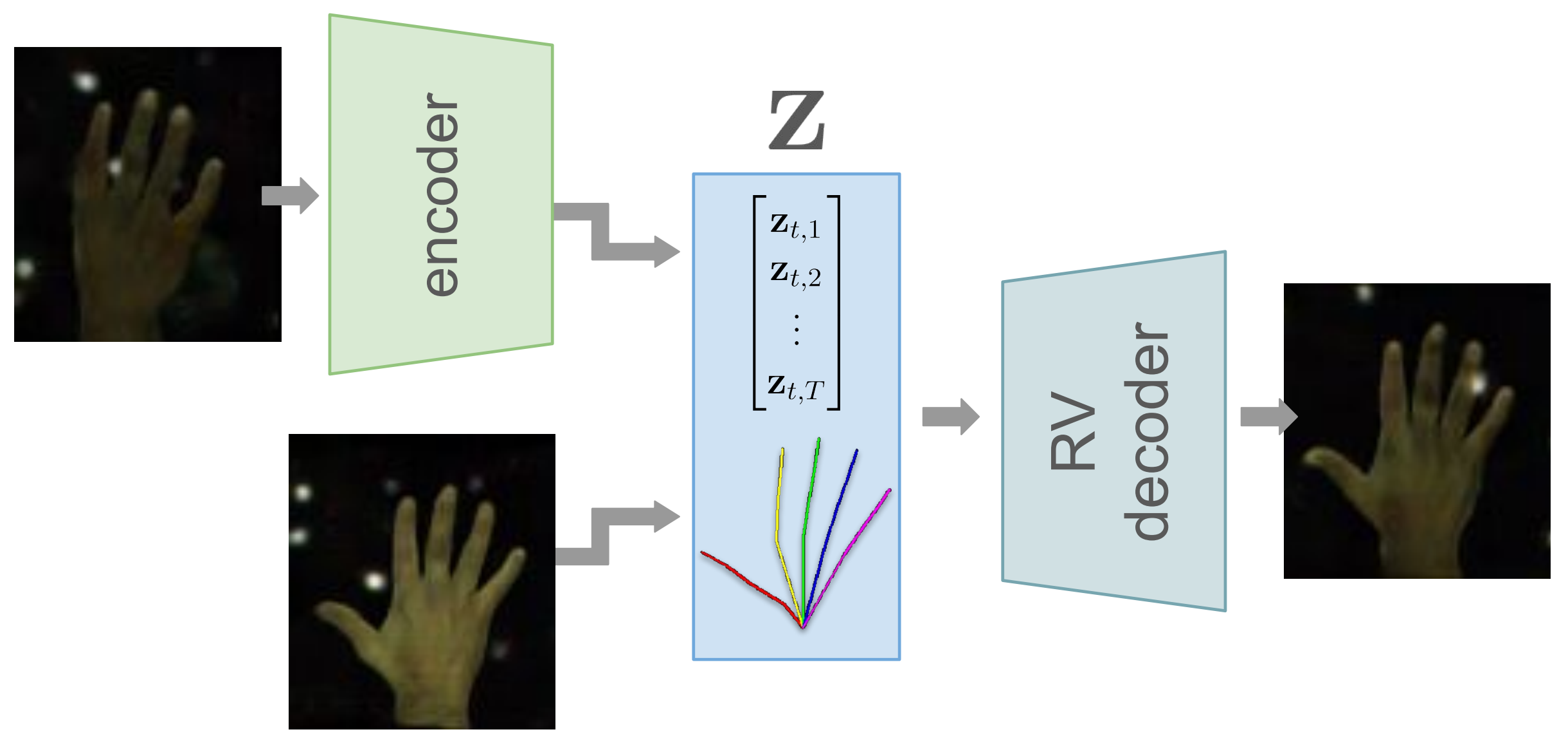
Method	Dataset	**FID ↓**
SRV-VAE	STB	26.84
SRV-VAE	InterHand2.6M	16.13
Soft-Intro-SRV-VAE	STB	14.62
Soft-Intro-SRV-VAE	InterHand2.6M	9.27

➢**Train poses** with **train appearances**

Method	Dataset	**FID ↓**
SRV-VAE	STB	25.27
SRV-VAE	InterHand2.6M	16.30
Soft-Intro-SRV-VAE	STB	11.07
Soft-Intro-SRV-VAE	InterHand2.6M	10.59

Utilization

➢Appearance transfer



➢Data augmentation for hand-specific tasks

Dataset	Original	Augmented
STB (pixel space)	11.74	**10.59**
InterHand2.6M (mm)	**11.51**	11.73