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], and Antonis Argyros[1,2]

[1] Computer Science Department, University of Crete, Heraklion, Greece
[2] Institute of Computer Science, FORTH, Heraklion, Greece
{nikodim,oikonom,gkarv,argyros}@ics.forth.gr

Abstract. We introduce a novel framework for generating photorealistic synthetic images of human hands conditioned to a precise pose annotation. We propose a supervised Random Variable Variational Autoencoder (SRV-VAE), a model that disentangles and encodes the appearance and pose of the hand into separate components of the latent space. Appearance, representing individual subject traits, is unsupervised. Hand pose is strictly supervised and yields control over the synthesis process. Leveraging the robust RV VAE variant, our architecture ensures stable training and accurate encoding of complex hand dynamics. Our model is capable of generating hand images of previously unseen hand poses for specific subjects. Experimental results indicate the model's efficacy in synthesizing realistic and varied hand images, holding significant promise for advancements in both academic research and practical applications such as data upsampling, where accurate hand pose and texture data is critical.

1 Introduction

The accurate estimation of hand pose, shape, and appearance from visual data, as well as the related problem of generating realistic hand images are challenging tasks with widespread applications in fields such as virtual reality, human-computer interaction, and robotics. A critical aspect of advancing these applications lies in the availability of large-scale, high-quality datasets that can be used to train robust machine learning models. While real-world data collection has made significant progress, particularly with the introduction of depth sensors and advances in deep learning techniques, the limitations in data quality, diversity, and annotation consistency still pose challenges to model development. To address these issues, synthetic data generation has emerged as a viable alternative, providing a scalable solution to augment real-world datasets and enhance model performance.

In recent years, generative models have garnered significant attention for their ability to create realistic data samples that closely resemble real-world data.

Techniques such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have demonstrated remarkable success across various domains, particularly in image generation. Despite these advances, generating realistic hand images remains a challenging task due to the complexity of hand anatomy and the need for precise control over hand poses. While GAN-based methods like GANerated Hands. [15] have made strides in synthesizing diverse hand poses, diffusion models have struggled with generating anatomically accurate hand images, often resulting in distorted outputs.

To overcome these challenges, this paper introduces a novel approach using a Supervised Random Variable Variational Autoencoder (SRV-VAE) for the generation of realistic hand images conditioned on specific hand poses. Our method builds upon the strengths of Random Variable VAE (RV-VAE) [17], which has been shown to effectively encode and take advantage of complex data representations by directly utilizing the entire distribution of the latent space, leading to improved performance in image generation tasks. By incorporating supervision into a part of the latent space for the pose and leaving the other part unsupervised for the appearance, our approach allows for fine-grained control over hand image generation. This is achieved by combining an appearance vector with a pose configuration, ensuring that the synthesized images are both realistic and pose-accurate even on unseen hand poses, as depicted in Fig. 1.

We present a comprehensive evaluation of our SRV-VAE framework, demonstrating its effectiveness in generating high-quality hand images across various poses. We compare our approach against existing methods, highlighting its advantages in terms of both qualitative and quantitative metrics. Additionally, we explore the utility of our generated images in enhancing the performance of hand pose estimation models, providing evidence of the broader applicability of our approach in augmenting hand image datasets.

Overall, our contributions are threefold: 1) We introduce SRV-VAE, a novel generative model that enables controlled and realistic hand image generation by disentangling hand pose and appearance in the latent space. 2) We demonstrate the effectiveness of SRV-VAE in generating high-quality hand images, even for previously unseen poses, by conducting extensive experiments on benchmark datasets. 3) We show that the generated images can be used to enhance the performance of hand pose estimation models, illustrating the practical value of our approach in real-world applications. By addressing the challenges of hand image generation with SRV-VAE, this work contributes to the ongoing efforts to improve hand pose estimation models and provides a robust framework for generating synthetic hand images with precise pose control.

2 Related Work

Research on visual hand pose, shape, and appearance estimation, as well as the related problem of image synthesis given this information, have progressed in strides in the recent years. Significant contributions that aided advance the field include the introduction of depth sensors [30], the deep learning revolu-

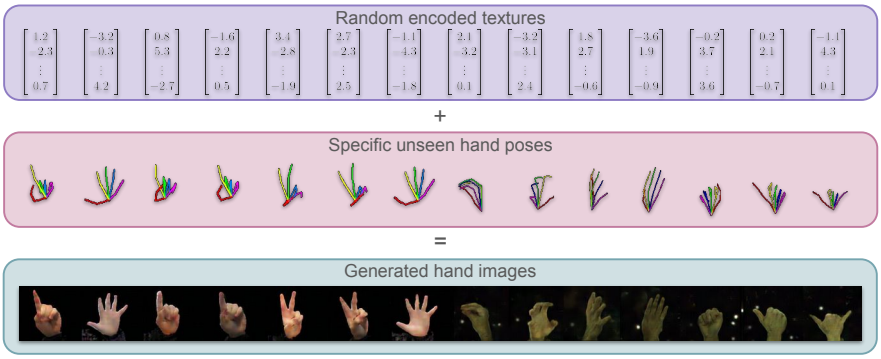

Fig. 1: By utilizing the normally distributed unsupervised latent texture space and the supervised hand pose space, our method is capable of generating realistic hand images even on unseen 2D/3D hand poses.

tion [11], and the availability of large, relevant, high quality datasets [3,15]. The state of the art today relies on increasingly accurate and high quality data to train better systems on [29]. This demand has led to bigger and better real-world datasets [14], but also to the introduction of synthetic data generation approaches to bootstrap or even complement training [12,13].

Generative models have drawn significant attention in recent years with their improving capabilities in creating new data samples that resemble real-world data. With the advent of deep learning, architectures such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have achieved remarkable success across various domains, including image generation, natural language processing, and music composition. For instance, GANs have been utilized to generate realistic images from textual descriptions [6,24]. Similarly, VAEs have shown proficiency in generating high-quality data by learning latent representations [9]. Specifically for image generation, one of the latest and most successful approaches involves diffusion models, which progressively denoise an initially random image to produce high-quality outputs [7,26]. The evolution of these models has led to remarkable instances, such as the DALL-E series, Stable Diffusion, and MidJourney, which generate high-quality, customizable images from textual descriptions, pushing the boundaries of generative models [25].

Despite this remarkable progress, diffusion models for images are notoriously bad at generating realistic hands, often producing distorted or anatomically incorrect results. Moreover, accurate hand pose annotation is crucial for generating synthetic data to be used for training machine learning models. Therefore, specialized techniques are still required to overcome these challenges and achieve controllable, precise, high quality hand image generation.

2.1 Hand generative models

The need for precise hand pose representation in synthesized, high-quality images has led to the development of dedicated generative models specifically tailored for hand images. "GANerated Hands" by Mueller *et al.* [15] is one of the early successful works on this topic. By leveraging Generative Adversarial Networks (GANs), the approach creates realistic hand images that can be used for training pose estimation systems. The method addresses the scarcity and limitations of existing hand datasets by generating a diverse array of hand poses and appearances that enhance the robustness of hand pose estimation models. While this method can produce realistic hand images, it requires a whole synthetic hand image instead of just one pose configuration. Moreover, this method makes synthetic hand images more realistic and does not add more variance to the texture/appearance of the generated hand images.

Further refining the capabilities of GANs in the domain of hand image synthesis, H-GAN, introduced by Oprea *et al.* [18], adopts a cyclic consistency approach [31] that improves the generation process and manages to focus separately on different aspects of the hand, such as preserving the pose of a given image, and altering the texture. While this approach allows for the creation of realistic looking images, it relies on 3D rendering to do so, whereas our proposed approach allows for the generation of new samples directly from real-world data by having control over the pose and the variance of the appearance.

Achieving realism in hand image synthesis requires careful consideration of lighting and illumination, a challenge that Chen *et al.* tackle in their work UR-Hand [2]. Their method enhances the realism of synthesized hand images by generating specific poses under varying lighting conditions. However, it maintains a consistent base texture for the hand, which limits the diversity in the appearance of the generated hands.

In the context of diffusion models, which have typically struggled with the complex structures of hands, the work of Yang *et al.* [27] introduces an innovative approach to improve the quality of synthesized hand images. By incorporating hand pose annotations and focusing on the accurate portrayal of hand anatomy, this method aims to overcome the common distortions and inaccuracies encountered in standard diffusion model outputs.

Another diffusion-based approach is HanDiffuser [16] which generates realistic images of hands in scenes described by a text prompt. This approach uses the pose of a hand which is extracted from an estimated full-body pose, as an intermediate supervision to generate the final image. The main input to this method is text and does not offer precise control over the poses.

Methods that are used with traditional rendering approaches are HTML by Fu *et al.* [23] and similirally Handy model [22]. Both methods disentagles the texture from the mesh of the hand, however are not suitable for direct image synthesis since both requires complex pipelines to estimate lighting conditions and the background.

2.2 Supervision in generative models

The incorporation of supervision techniques into generative models has proven to be a powerful strategy for enhancing model performance and reliability. One of the notable advancements in this area, for the related problem of controlling the body pose of a depicted human, is ControlNet by Zhang *et al.* [29], which integrates supervised learning to guide diffusion models in the image generation process. The resulting method is suitable for tasks requiring accurate control of human poses. While this method can produce realistic images that contain hands given a 2D pose, it requires the whole body pose and struggles to generate images that are focused on hands explicitly.

Supervised approaches have also been extended to autoencoders and Variational Autoencoders (VAEs). For instance, the work by Le *et al.* [10] introduces supervision into autoencoders to enhance learning efficiency and output consistency. This can be particularly beneficial in applications such as medical image analysis. In VAEs, Berkhahn *et al.* [1] demonstrate how supervised learning can be utilized to enforce specific properties in the generated images, further enhancing the utility of VAEs in complex image generation tasks like hand pose estimation. Additionally, the integration of supervision techniques has proven effective in anomaly detection, as shown by Kawachi *et al.* [8]. Their approach uses supervised learning to refine the model's ability to identify and differentiate normal from anomalous patterns, which is crucial for ensuring the quality and usability of generated datasets in training other models.

Overall, the current state-of-the-art can achieve high quality hand images of a given pose using diffusion models, however at a high computational cost. Faster approaches such as GANs and VAEs have their own limitations, such as requiring an input image of the target hand pose, and poor generalization to unseen poses and variety of appearances. In this work, we present an approach that can bridge these gaps, achieving fast, high quality hand image generation, including previously unseen hand poses.

3 Methodology

In this work we present a novel approach that employs Supervised Random Variable VAE (SRV-VAE) for the synthesis of realistic hand images given a known pose. SRV-VAE facilitates the disentanglement of hand pose and arbitrary appearance vectors, crucial for conditional generation, allowing control over the generation process. An overview of our approach is depicted in Fig. 2. By leveraging the stable training and precise encoding capabilities of RV-VAE [17] (Sec. 3.1), we establish a partially supervised latent space for hand poses by employing conditioning modifications in the VAE architecture (Sec. 3.2). This way, hand pose, and appearance features are effectively disentangled within the latent space, producing a visual combination of the two during the forward pass of SRV-VAE (Sec. 3.3). The resulting encoder provides an estimation of the input hand pose and encodes the appearance information separately from RGB

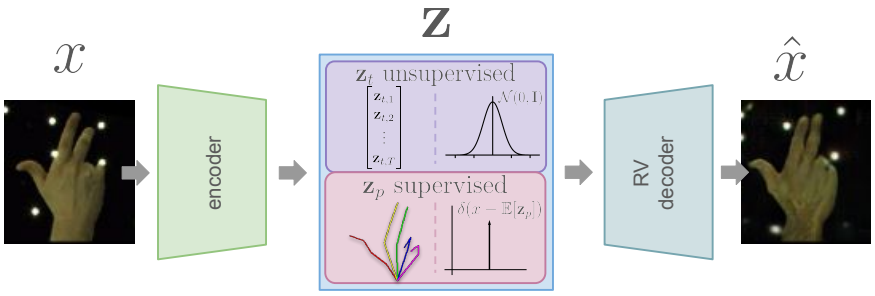

Fig. 2: The proposed SRV-VAE architecture, for an RGB hand image input x, disentangles the latent space into the unsupervised random variable $\mathbf{z}_t$, and the supervised random variable $\mathbf{z}_p$. The $\mathbf{z}_t$ random variable depicts the encoded texture vector and follows a standard normal distribution, while the $\mathbf{z}_p$ random variable depicts the estimated hand pose and follows a δ distribution. By leveraging the capabilities of the RV-aware architecture we forward these distributions directly to the decoder for reconstructing the input RGB hand image.

hand images, while the decoder generates realistic hand images based on specific poses and arbitrary appearance vectors.

3.1 RV-VAE

The formation of the latent space and the generative capabilities of VAEs is attributed to the training procedure of optimizing the ELBO loss as defined by Kingma and Welling in [9] and shown in Eq. 1. We follow the standard notation as used in that work [9].

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]. \tag{1}$$

The symbols used denote: $q_\phi(\mathbf{z}|\mathbf{x})$ is the approximate posterior distribution over the latent variable $\mathbf{z}$ given the data $\mathbf{x}$, parameterized by ϕ and typically modeled using a neural network, the encoder. Similarly, $\log p_\theta(\mathbf{x}|\mathbf{z})$ is the log-likelihood of the data $\mathbf{x}$ given the latent variable $\mathbf{z}$, parameterized by θ and typically modeled as another neural network, the decoder. Additionally, $p_\theta(\mathbf{z})$ denotes the prior distribution over the latent variable $\mathbf{z}$, often chosen to be a simple distribution like a standard normal distribution $\mathcal{N}(0, 1)$. The term $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}$ denotes the expectation of the log-likelihood with respect to the approximate posterior distribution over $\mathbf{z}$. Finally, the term D_{KL} denotes the Kullback-Leibler divergence between the two distributions, the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the chosen prior $p_\theta(\mathbf{z})$.

Since the second term in this loss formulation, the expectation $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}$ is practically intractable, the authors suggest forming Monte Carlo estimates of it, with the final estimation becoming:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) \simeq \tilde{\mathcal{L}}(\theta, \phi; \mathbf{x}) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) + \frac{1}{L}\sum_{l=1}^{L}(\log p_\theta(\mathbf{x}|\mathbf{z}^{(l)})), \tag{2}$$

where L denotes the number of samples drawn, as implemented by the reparameterization trick.

Nicodemou *et al.* [17] showed that this sampling can be avoided by using special differentiable Random Variable operations inside the decoder's architecture, and in this way utilizing the whole distribution of $q_\phi(\mathbf{z}|\mathbf{x})$ which describes the encoder's output. This modification has been shown to improve the overall performance of VAEs in terms of reconstruction and image generation without hindering the convergence rate.

Empirical evidence (Sec. 4.4) shows that RV-VAEs manage to better disentangle complex data representations, such as those of hands, particularly when using supervised encoded attributes, such as hand poses. This is ascribed to the fact that parts of the latent space (supervised or not) are utilized completely by the network as RVs and prevent any data loss since there are no sampling procedures. This results in the accurate and distinct formulation of encoded regions of the latent space.

3.2 Supervised RV-VAE

The beneficial usage of RV-VAE has been reported in [17] for image generation on multiple datasets. However, following this methodology gives no control over any specific attributes we would like to impose in the generation process. Any specific image creation given a requested attribute would require a search in the "opaque" latent space. Therefore we need to enforce some form of structure in the latent space. To achieve this, a natural choice is the language of conditional probability distributions, already used since the original formulation of VAEs. In our case, our goal is to approximate the conditional probability of a hand image given a specific hand pose, disentangled from the rest of the latent space.

The conditioning of image generation is achieved by supervising a subset of the latent space within the RV-VAE training procedure. By incorporating supervision into specific dimensions of the latent space, we aim to impart control over specific aspects of image generation. This novel strategy enables the generation of images conditioned not only on random noise but also on structured latent representations.

Specifically, the whole latent space $S \subseteq \mathbb{R}^n$ (for suitable dimensionality n) is divided into two sub-spaces, the new supervised and the regular unsupervised sub-space. This is achieved by the output of the encoder in the RV-VAE architecture with a modification of the encoder's final layer. Specifically, for a hand pose p with D spatial dimensions and K keypoints, and a latent texture vector of size T, the encoder outputs the parameters of a latent random variable $\mathbf{z} \in \mathbb{R}^{(D\times K+T)\times 2}$ with the last dimension being the two distribution parameters, mean and variance. The random variable $\mathbf{z}$ is the concatenation of the random variable $\mathbf{z}_p \in \mathbb{R}^{D\times K}$ and $\mathbf{z}_t \in \mathbb{R}^T$ that depict the hand pose and the encoded latent texture space of the input, respectively.

The general form of the ELBO loss in Eq. 1 is modified to incorporate the new conditionality of the latent space, and is given by:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = - D_{KL}(q_\phi(\mathbf{z}_t|\mathbf{x})||p(\mathbf{z}_t)) + \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x}), q_\phi(\mathbf{z}_p|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathbb{E}_{p(\mathbf{z}_p|\mathbf{x})}[\log q_\phi(\mathbf{z}_p|\mathbf{x})]. \tag{3}$$

In Eq. 3, $q_\phi(\mathbf{z}_t|\mathbf{x})$ is the encoder's first output that we want to match to the prior distribution $p(\mathbf{z}_t)$ over the latent variable $\mathbf{z}_t$ (which is the standard normal distribution $\mathcal{N}(0, \mathbf{I})$). $p_\theta(\mathbf{x}|\mathbf{z})$ is the decoder's output, that is, the reconstruction of $\mathbf{x}$ given $\mathbf{z}$ where $\mathbf{z}$ is the concatenation of both random variables $\mathbf{z}_t$, $\mathbf{z}_p$. Finally, $q_\phi(\mathbf{z}_p|\mathbf{x})$ is the encoder's second output representing the regressed hand pose. This is compared to the true posterior $p(\mathbf{z}_p|\mathbf{x})$, that is, the known hand pose $\mathbf{x}$.

The encoder outputs the parameterization of a distribution, specifically its first two moments. For a training set of hand images, the ground truth hand poses can be described by a degenerate distribution $P(X = \mathbf{p}) = 1$, separately for each hand pose $\mathbf{p}$ as they can be considered independent constant random variables with a probability density function described by Dirac delta function $\delta(x - \mathbf{p})$. In this paper the goal is hand image generation conditioned on a given hand pose. Therefore, for an input image $\mathbf{x}$ with associated hand pose $\mathbf{p}$, the optimization of $\mathbb{E}_{p(\mathbf{z}_p|\mathbf{x})}[\log q_\phi(\mathbf{z}_p|\mathbf{x})]$ term from Eq. 3 is equivalent to minimizing the mean $\frac{1}{K}\sum_{i=1}^{K}(\mathbf{p}_i - \hat{\mathbf{p}}_i)^2$ and the variance $\frac{1}{K}\sum_{i=1}^{K}(\text{var}[\mathbf{z}_{p_i}])^2$ of the Encoder's output, where $\mathbf{p}$ and $\hat{\mathbf{p}} = \mathbb{E}[\mathbf{z}_p]$ are the ground truth pose with K keypoints and the estimated (by the encoder) hand pose, respectively.

Since the encoder outputs the parameters of distributions, by utilizing the advantages of RV-VAEs we can forward these distributions (of pose and appearance) directly to the decoder. This is possible since the modules inside the decoder are designed to operate on random variables instead of samples (like the reparameterization trick would provide in regular VAEs).

3.3 Forward pass of SRV-VAE

Given the formulation described in Sec. 3.2, during inference, for a test sample $\mathbf{x}$, the encoder will output its estimated hand pose in the form of $\mathbb{E}[\mathbf{z}_p]$.

While the encoder's estimation of hand poses from RGB images might be straightforward, the generation of new images requires further elaboration due to the treatment of the latent data as random variables. Specifically, in order to create an RGB hand image we require a hand pose $\mathbf{p}$ that we desire to generate and a random vector texture encodings $\mathbf{z}_t$. By concatenating the flattened vector of $\mathbf{p}$ with the texture vector $\mathbf{z}_t$ we create the latent vector $\mathbf{z}$. The decoder then takes this representation of a latent distribution, and in contrast to regular VAE architectures [9], it does not perform a reparameterization trick. Instead, using the RV-VAE approach [17], the two distributions are propagated throughout the layers of the decoder toward the output, with one being the distribution of encoded hand texture $\mathcal{N}(\mathbb{E}[\mathbf{z}_t], \text{var}[\mathbf{z}_t])$, and the other of the hand pose $\delta(x - \mathbb{E}[\mathbf{z}_p])$. The output will be a generated hand image of $\mathbb{E}[\mathbf{z}_p]$ pose and visualized with the respective texture appearance.

4 Experiments

We conducted several experiments over a couple of models and datasets to evaluate and assess the performance of the proposed SRV-VAE framework. Specifically, we created variations of a regular VAE architecture [9] and of one based on Soft-Intro-VAE [4], a state-of-the-art approach in generative VAE architectures. These variations were modified to become RV-aware models based on the approach by Nicodemou *et al.* [17], and further modified to incorporate latent space conditionality, as described in this work. Both model architectures were trained on two datasets, the Stereo Hand Pose Benchmark (STB) [28] with 2D keypoint hand poses and the InterHand2.6M [14] with 3D keypoint hand poses. Experiments focused on highlighting the quality of hand images conditioned on specific poses (Sec. 4.2) as well as the quantitative assessment of the generative capabilities (Sec. 4.3). Since this work depends on the beneficial usage of RV-VAEs, we investigated the contribution that RV-aware models provide to the conditionality problem we tackle in this work, compared to regular S-VAE architectures (Sec. 4.4). The byproduct of generating images from disentangled latent space can be utilized as appearance transfer between poses (Sec. 4.5) or up-sampling sparse hand datasets (Sec. 4.6), both strengthen the motivation of this work. All experiments were conducted on the Stereo Hand Pose Benchmark (STB) [28] and the InterHand2.6M [14] datasets.

4.1 Implementation Details

All models and architectures were implemented using the PyTorch [20] library. From the STB dataset, we used sequences labeled as "Random" for training (9k samples) and sequences labeled as "Counting" for testing (9k samples). For hand pose ground truth annotation, we extracted and used 2D hand keypoints. From the InterHand2.6M, we used the train/test split of single right hand images defined by the dataset. From both sequences we removed frames and poses with occlusions or erroneous visual annotations, resulting in 20k training and testing samples. For this dataset, we used 3D hand keypoints.

4.2 Qualitative Results

To illustrate the proposed method's generative capabilities we report some qualitative results for all trained methods on both datasets. Specifically, Fig. 3 illustrates the generative results on both datasets from SRV-VAE. Furthermore, Figs. 4 and 5 show the generative results from the Soft-Intro-SRV-VAE architecture. All images were generated by using unseen test hand poses and by changing their appearance randomly, by concatenating each time a different random texture vector with the test hand pose. We can observe the high quality of generated images resulting from the Soft-Intro-SRV-VAE. The Soft-Intro-RV-VAE's architecture and training procedure yields a higher quality of images compared to the regular RV-VAE architecture. We can also observe some slight changes in the generated poses (specifically in the last two columns of Fig. 4) when iterating

(a) STB dataset **(b)** InterHand2.6M

Fig. 3: Generated hand images using the SRV-VAE model on the two datasets. Each column has fixed test (unseen) poses, and each row changes the random appearance vector.

through random appearance vectors. This can be attributed to the fact that the specific training set (STB) consists of fewer samples than the InterHand2.6M, making the disentanglement of pose and appearance more challenging.

4.3 Quantitative Results

A commonly used and reliable method to quantitatively assess the quality of generated samples of a generative model based on a training dataset is by using the Fréchet Inception Distance (FID) metric [4, 17, 19]. This metric measures the distance between the distributions of real and generated images, providing a quantitative assessment of how similar the generated images are to the real ones.

Given the nature of the problem this work aims to tackle, the generative process does not depend only on the training set but also on a test set, as generated images are conditioned on the unseen poses of that test set. Therefore, the comparison between the distributions of training and generated images is not indicative, since by design, we want them to differ. For that reason, we need to consider two different comparisons: (a) between the distributions of training images and generated images conditioned on poses seen in the training set, and (b) between the distributions of test images and generated images conditioned on unseen poses from the test set. We want those differences not to be far apart, showing that the quality does not change drastically when generated from different distributions.

In Tab. 1 we report the FIDs that measure the distance between the distribution of training set images and the distribution of generated images from poses of the same training set. Respectively, in Tab. 2 we present the FIDs that measure the distance between the distribution of test set images and the distribution of generated images from unseen poses of the test set. As expected, differences when comparing the generated images from seen and unseen poses are very small, indicating a good quality of hand images.

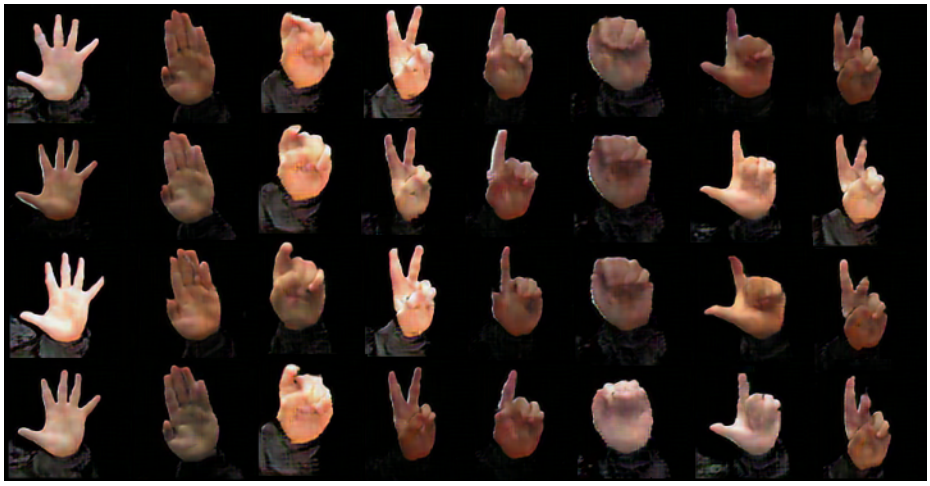

Fig. 4: Generated hand images using the Soft-Intro-SRV-VAE model on STB dataset. Each column has fixed test (unseen) poses, and each row changes the random appearance vector.

Table 1: FIDs values that measure distances for generated images between real training set distribution and generated (known training poses) distribution for different combinations of datasets and methods.

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

4.4 SRV-VAE vs regular S-VAE

To make a fair comparison between SRV-VAE and regular S-VAE that specialize in conditioning the latent space as stated by our problem, we are required to sample from two sub-spaces (the pose and texture) for the regular S-VAE. This is crucial for the training of the S-VAE network, while the SRV-VAE takes all distributions as they are (Sec. 3.2).

We trained both networks, SRV-VAE and S-VAE, on the STB dataset and validated the comparison on the test set as described in Sec. 4.3. Table 3 reports the FIDs on the regular supervised VAE compared to the previously reported FIDs of the RV variant. We observe that the RV modules in the network contribute significantly towards the generative capabilities of the method. This improvement can be attributed to the fact that SRV-VAE does not depend on any sampling during training, whereas the regular S-VAE requires sampling from two spaces. This double sampling introduces even more uncertainty into the training pipeline.

Fig. 5: Generated hand images using the Soft-Intro-SRV-VAE model on InterHand2.6M dataset. Each column has fixed test (unseen) poses, and each row changes the random appearance vector.

Table 2: FIDs values that measure distances for generated images between real test set distribution and generated (unseen test poses) distribution for different combinations of datasets and methods.

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

4.5 Appearance Transfer

The formulation we use in this work disentangles the appearance and pose of an image depicting a hand as explained in Sec. 3.2. The disentanglement is enforced by the encoder of the network, which outputs separately: an estimation of the encoded texture, random variable $\mathbf{z}_t$, and an estimation of the hand pose of the input, random variable $\mathbf{z}_p$. This implies that a trained encoder can yield an appearance estimation of an input image that can be "enforced" on a different pose to generate new images as seen in Fig. 6.

4.6 Quantitative Evaluation on Downstream Problems

Given the nature of the proposed approach, a meaningful question to ask is whether it can be used to improve the quality of existing datasets, by performing a domain-aware data augmentation. Specifically, given a hand pose dataset,

Table 3: FIDs values that measure distances for generated images between real test set distribution and generated (unseen test poses) distribution for different combinations of datasets and methods.

Method	Generation Type	**FID** ↓
STB on S-VAE	Train (seen) poses	26.41
STB on S-VAE	Test (unseen) poses	29.72
STB on SRV-VAE	Train (seen) poses	**25.27**
STB on SRV-VAE	Test (unseen) poses	**26.84**

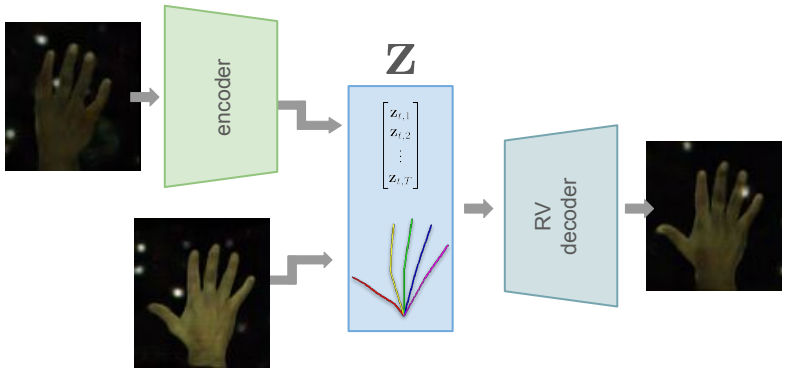

Fig. 6: The trained encoder can be used to extract the appearance of an input image that can be transferred to a different pose via the decoder.

one can use the proposed approach to generate more training samples, by combining hand poses from one subject with the appearance of another, effectively "upsampling" the existing dataset into a denser one. Such an approach would enrich the existing dataset, potentially improving the accuracy of a keypoint detector trained on the resulting augmented data. To assess this hypothesis, we implemented a hand keypoint detector network based on the ViT architecture [5]. Specifically, using ViT as the backbone, we added a head for the prediction of the hand joint positions, inspired by [21]. The resulting network consists of 10M parameters.

We trained this detector on two subsets of the datasets we used, STB and InterHand2.6M. Each was augmented, and we compared the performance with and without the augmented data. Specifically, we used subsets of the original InterHand2.6M and STB datasets consisting of 4000 and 1800 images, respectively. These datasets were augmented with 8000 and 3600 images. We also selected a test set for each, consisting of 4100 and 1800 images, respectively. To ensure a fair comparison, the networks trained on the original datasets were trained for twice as many epochs as those with augmented datasets, matching the doubled amount of training data in the latter.

Table 4: Mean per Joint Positional Error (MPJPE) comparing the performance of the implemented keypoint estimator on original and augmented versions of the two datasets we experimented on.

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

As shown in Tab. 4, the augmentation is always at least non-disrupting, and in the case of one of the datasets, it helps significantly reduce the resulting estimation error. This is attributed to the diversity difference between the two datasets: STB has only one subject gesturing a limited range of hand gestures, whereas InterHand2.6M is more diverse, both in subjects, and in the performed gestures. These results show that our approach is particularly beneficial as a dataset augmentation tool for the case of small or non-diverse datasets, while not hurting the final performance in other cases.

5 Conclusions

In this paper, we introduced SRV-VAE, a novel supervised variational autoencoder framework designed to generate realistic hand images conditioned on specific hand poses. Our approach effectively addresses the challenges associated with hand image generation, particularly the need for precise control over pose and the synthesis of anatomically accurate hand images. By leveraging the strengths of the RV-VAE architecture and incorporating supervision into the latent space, SRV-VAE enables the disentanglement of pose and appearance, resulting in high-quality and diverse hand image outputs. Our experiments show that SRV-VAE produces visually convincing images even from unseen poses. This outcome can be beneficial in augmenting datasets with the task of enhancing the performance of hand pose estimation models. Future work includes expanding the SRV-VAE to disentangle the latent space even further, specifically the appearance domain into separate subdomains (shape, texture, illumination, etc.), and investigating the use of SRV-VAE in other aspects of real-world data.

Acknowledgements

This work was co-funded by: (a) the European Union (EU - HE Magician – Grant Agreement 101120731), (b) the Hellenic Foundation for Research and Innovation (HFRI) under the “1st Call for HFRI Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment”, project I.C.Humans, no 91, (c) the internal ICS project "Large-scale, Diverse 3D Modeling of Human Hands" - DiveHands. The authors also gratefully acknowledge the support for this research from the VMware University Research Fund (VMURF).

References

1. Berkhahn, F., Keys, R., Ouertani, W., Shetty, N., Geißler, D.: Augmenting variational autoencoders with sparse labels: A unified framework for unsupervised, semi-(un) supervised, and supervised learning. arXiv preprint arXiv:1908.03015 (2019) 5
2. Chen, Z., Moon, G., Guo, K., Cao, C., Pidhorskyi, S., Simon, T., Joshi, R., Dong, Y., Xu, Y., Pires, B., Wen, H., Evans, L., Peng, B., Buffalini, J., Trimble, A., McPhail, K., Schoeller, M., Yu, S.I., Romero, J., Zollhöfer, M., Sheikh, Y., Liu, Z., Saito, S.: URhand: Universal relightable hands. In: CVPR (2024) 4
3. Christian Zimmermann, Duygu Ceylan, J.Y.B.R.M.A., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: IEEE International Conference on Computer Vision (ICCV) (2019), "https://lmb.informatik.uni-freiburg.de/projects/freihand/" 3
4. Daniel, T., Tamar, A.: Soft-introvae: Analyzing and improving the introspective variational autoencoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4391–4400 (June 2021) 9, 10
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020) 13
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 3
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3
8. Kawachi, Y., Koizumi, Y., Harada, N.: Complementary set variational autoencoder for supervised anomaly detection. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2366–2370. IEEE (2018) 5
9. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (2014) 3, 6, 8, 9
10. Le, L., Patterson, A., White, M.: Supervised autoencoders: Improving generalization performance with unsupervised regularizers. Advances in neural information processing systems **31** (2018) 5
11. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015) 3
12. Lee, J., Kim, J., Kim, S.H., Choi, S.I.: Enhancing 3d hand pose estimation using shaf: synthetic hand dataset including a forearm. Applied Intelligence pp. 1–14 (2024) 3
13. Li, L., Tian, L., Zhang, X., Wang, Q., Zhang, B., Bo, L., Liu, M., Chen, C.: Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20395–20405 (2023) 3
14. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: European Conference on Computer Vision (ECCV) (2020) 3, 9

15. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: Ganerated hands for real-time 3d hand tracking from monocular rgb. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 49–59 (2018) 2, 3, 4
16. Narasimhaswamy, S., Bhattacharya, U., Chen, X., Dasgupta, I., Mitra, S., Hoai, M.: Handiffuser: Text-to-image generation with realistic hand appearances. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 4
17. Nicodemou, V.C., Oikonomidis, I., Argyros, A.: Rv-vae: Integrating random variable algebra into variational autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 196–205 (2023) 2, 5, 7, 8, 9, 10
18. Oprea, S., Karvounas, G., Martinez-Gonzalez, P., Kyriazis, N., Orts-Escolano, S., Oikonomidis, I., Garcia-Garcia, A., Tsoli, A., Garcia-Rodriguez, J., Argyros, A.: H-gan: the power of gans in your hands. In: 2021 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2021) 4
19. Parmar, G., Li, D., Lee, K., Tu, Z.: Dual contradistinctive generative autoencoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 823–832 (June 2021) 10
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019) 9
21. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9826–9836 (2024) 13
22. Potamias, R.A., Ploumpis, S., Moschoglou, S., Triantafyllou, V., Zafeiriou, S.: Handy: Towards a high fidelity 3d hand shape and appearance model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4670–4680 (June 2023) 4
23. Qian, N., Wang, J., Mueller, F., Bernard, F., Golyanik, V., Theobalt, C.: HTML: A Parametric Hand Texture Model for 3D Hand Reconstruction and Personalization. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer (2020) 4
24. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015) 3
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 3
26. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019) 3
27. Yang, Y., Gandhi, A.N., Turk, G.: Annotated hands for generative models. arXiv preprint arXiv:2401.15075 (2024) 4
28. Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., Yang, Q.: 3D Hand Pose Tracking and Estimation Using Stereo Matching. arXiv:1610.07214 (2016) 9

29. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 3, 5
30. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE multimedia **19**(2), 4–10 (2012) 2
31. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017) 4