Deep Models for Generation of 3D Characters and Animations from Sketches

by

Alican Akman

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computer Science and Engineering

KOÇ ÜNİVERSİTESİ

August 6, 2020

Deep Models for Generation of 3D Characters and Animations from Sketches

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Alican Akman

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. T. Metin Sezgin (Advisor)

_____

Assoc. Prof. Yusuf Sahillioğlu

_____

Prof. Ergun Akleman

Date: _____

To my grandfather who lived and worked for his grandchildren

ABSTRACT

Deep Models for Generation of 3D Characters and Animations from Sketches

Alican Akman

Master of Science in Computer Science and Engineering

August 6, 2020

Generating 3D models from 2D images or sketches is a widely studied important problem in computer graphics. We describe the first method to generate a 3D human model from a single sketched stick figure. In contrast to the existing human modeling techniques, our method requires neither a statistical body shape model nor a rigged 3D character model. We exploit Variational Autoencoders to develop a novel framework capable of transitioning from a simple 2D stick figure sketch, to a corresponding 3D human model. Our network learns the mapping between the input sketch and the output 3D model. Furthermore, our model learns the embedding space around these models. We demonstrate that our network can generate not only 3D models, but also 3D animations through interpolation and extrapolation in the learned embedding space. In addition to 3D human models, we produce 3D horse models in order to show the generalization ability of our framework. Extensive experiments show that our model learns to generate reasonable 3D models and animations.

ÖZETÇE

Basit Çizimleri Kullanarak 3 Boyutlu Model ve Animasyon Üretimi
Alican Akman
Bilgisayar Bilimi ve Mühendisliği, Yüksek Lisans
1 Ocak 2019

2 boyutlu imajlardan ve çizimlerden 3 boyutlu modeller üretmek, bilgisayar grafikleri alanında genişçe çalışılan önemli bir problemdir. Biz bu tezde, tek bir çizilmiş çubuk şeklinden 3 boyutlu bir insan modeli üreten ilk metotu tanımlıyoruz. Mevcut olan insan modelleme tekniklerine karşın bizim metotumuz, ne bir istatiksel vücut şekli modeline ne de bir giydirilmiş 3 boyutlu karakter modeline ihtiyaç duymaktadır. Biz bu çalışmada basit bir 2 boyutlu çubuk şekil çiziminden, bu çizimle uyumlu olan bir 3 boyutlu insan modeline dönüşüm kabiliyetine sahip yeni bir sistem geliştirmek için varyasyonel otokodlayıcılardan faydalanırız. Bizim makine öğrenmesi modelimiz, girdi çizim ile çıktı 3 boyutlu model arasındaki eşleştirmeyi öğrenmektedir. Bunun yanında modelimiz, bu modeller etrafında kurulan gömme uzayı da öğrenmektedir. Biz sistemimizin 3 boyutlu modeller üretmenin haricinde, bu gömme uzaydaki enterpolasyon ve ekstrapolasyonlar ile birlikte 3 boyutlu animasyonlar da üretebildiğini göstermekteyiz. 3 boyutlu insan modellerine ek olarak, sistemimizin genelleme yeteneğini göstermek için 3 boyutlu at modelleri de üretiriz. Kapsamlı deneyler, modelimizin mantıksal 3 boyutlu modeller ve animasyonlar ürettiğini göstermektedir.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Prof. Metin Sezgin for his vast support during my masters. He has propelled me to become a strong researcher by sharing his insight, creative perspective, amazing ideas, and experience. I am grateful that I have had the chance to work with Prof. Yusuf Sahillioğlu on my thesis subject and our published papers. His inspiring advice led me in the right direction for my research. In addition to his amazing support on my academic skills, he showed me how a new generation academician should be with his positive and motivational attitude. I also would like to thank Prof. Ergun Akleman for being in my thesis committee and the other professors from whom I have taken marvelous courses. As a person who believes that touching other people's lives is one of the most divine acts in this life, I will always be indebted to my professors who have contributed to my development.

I was a member of the Intelligent User Interfaces (IUI) lab at Koc University. We are an incredible team who love helping each other while performing our individual research. I would like to thank my dear friends Berker, Ozan, Zana, Kurmanbek, Dogancan, Elif, and Alpay. They all have a significant role in my academic study and I feel lucky to have met these awesome people and to have spent priceless time with them.

My deepest thanks go to my family for always standing by me throughout my life. I have always felt strong when seeing their essential support and love behind me. I could not have come to these days and could not be writing this thesis without them. They gave me both the personal and financial independence that enabled me to make decisions for my future without any concern. I will always try my best to deserve their support and make them proud.

I would like to thank my dearest friends, Ahmet Serdar, Furkan, Hasan Burak,

İlker, Nafi, Ömer Burak, and Yusuf. I see all of these people as part of my family. We have shared great moments and supported each other in good and bad times. Last but most important, I am grateful to my dearest, Hande, for standing by with me at every moment of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

2D	Two-dimensional
3D	Three-dimensional
AE	Autoencoder
VAE	Variational Autoencoder
ReLU	Rectified Linear Activation Unit
KL	Kullback–Leibler
Std	Standard Deviation
CPU	Central Processing Unit
GPU	Graphics Processing Unit
GHz	Gigahertz

Chapter 1

INTRODUCTION

3D content is still not as big as image and video data. One of the main reasons of this lack of abundance is the labor going into the creation process. Despite the increasing number of talented artists and automated tools, it is obviously not as simple and quick as hitting a record button on the phone.

3D content is, on the other hand, as important as the image and video data since it is used in many useful pipelines ranging from 3D printing to 3D gaming and filming.

With these considerations in mind, we aim to make the important 3D content creation task simpler and faster. To this end, we train a neural network over 2D stick figure and corresponding 3D model pairs. Utilization of easy-to-sketch and sufficiently-expressive 2D stick figures is a unique feature of our system that makes our system work properly even with a moderate amount of training, e.g., 72 distinct poses of a human model and 8 distinct poses of a horse model are used. We focus on human models as they come forward in 3D applications and extend our generation ability on horse models.

Given 2D stick figure sketches, our algorithm is able to produce visually appealing 3D point cloud models without requiring any other input such as a rigged template model. After an easy and seamless tweaking in the network, the system is also capable of producing dynamic 3D models, i.e., animations, between source and target stick figures as shown in Figure 1.1.

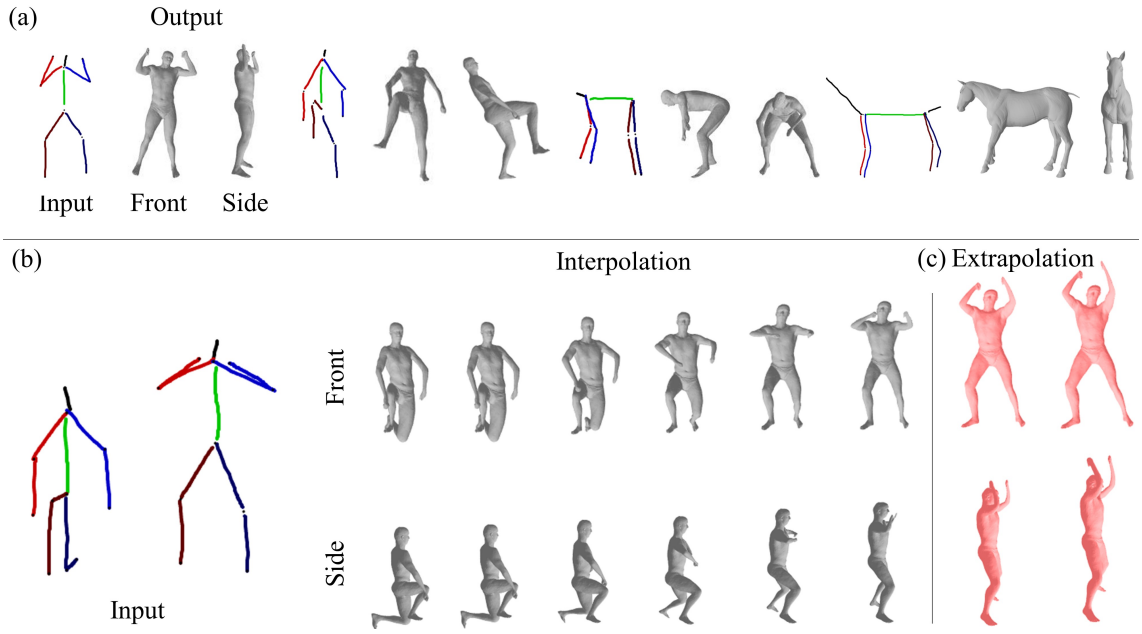

Figure 1.1: Our framework is capable of performing three tasks. (a) It can generate 3D models from given 2D stick figure sketches. (b) It can generate dynamic 3D models, i.e., animations, between given source and target stick figures. (c) It can further extrapolate the produced 3D model sequence by using the learned interpolation vector.

Chapter 2

RELATED WORK

Thanks to their natural expressiveness power, sketches are common modes for interaction for various graphics applications [Sezgin et al., 2006, Olsen et al., 2009].

The majority of sketch-based 3D human modeling methods deal with re-posing a rigged input model under the guidance of user sketches. [Guay et al., 2013] performs this action by transforming imaginary lines running down a character's major bone chains, whereas [Öztireli et al., 2013] and [Hahn et al., 2015] propose incremental schemes that pose characters one limb at a time. 2D stick figures to pose characters benefit from user annotations [Davis et al., 2003], specific priors [Lin et al., 2012], and database queries [Wei et al., 2011, Choi et al., 2012]. Bessmeltsev et al. [Bessmeltsev et al., 2016] claim that ambiguity problems of all these methods can be alleviated by contour-based gesture drawing. Deep regression network of [Han et al., 2017] utilizes contour drawing to allow face creation in minutes. Another system which takes one or more contour drawings as its input uses deep convolutional neural networks to create variety of 3D shapes [Delanoy et al., 2018]. We avoid the need for a rigged model in input specification and merely require a user sketch, which, when fed into our network, produces the 3D model quickly in the specified pose. For example, as the network is trained with the SCAPE models [Anguelov et al., 2005], our resulting 3D shape looks like the SCAPE actor, i.e., a 30 year-old fit man.

There also exist sketch-based modeling methods for other specific objects such as hairs [Fu et al., 2007, Seki and Igarashi, 2017] and plants [Anastacio et al., 2006], as well as general-purpose methods that are not restricted to a particular object class. These generic methods, consequently, may not perform as accurately as their object-specific counterparts for those objects but still demonstrate impressive results. 3D free-form design by the pioneer Teddy model [Igarashi et al., 1999] is improved in FiberMesh [Nealen et al., 2007] and SmoothSketch [Karpenko and Hughes, 2006]

by removing the potential cusps and T-junctions with the addition of features such as topological shape reconstruction and hidden contour completion. The recent SymmSketch system [Miao et al., 2015] exploits symmetry assumption to generate symmetric 3D free-form models from 2D sketches. In order to increase quality in generating 3D models, [Smirnov et al., 2019] focuses piecewise-smooth man-made shapes. Their deep neural network-based system infers a set of parametric surfaces that realize the drawing in 3D. Other solutions to the sketch-based generic 3D model creation problem depend on image guidance [Tsang et al., 2004, Yang et al., 2005], snapping one of the predefined primitives to the sketch by fitting its projection to the sketch lines [Shtof et al., 2013], and controlled curvature variation patterns [Li et al., 2017].

3D model generation and editing have been extended to 3D scenes as well. Dating back to 1996 [Zeleznik et al., 1996], this line of works generally index 3D model repositories by their 2D projections from multiple views and retrieve the elements that best match the 2D sketch query [Shin and Igarashi, 2007, Lee and Funkhouser, 2008]. Xu et al. [Xu et al., 2013] extend this idea further by jointly processing the sketched objects, e.g., while a single computer mouse sketch is not easy to recognize, other input sketches such as computer keyboard may provide useful cues. Sketch-based user interfaces arise in 2D image generation as well [Chen et al., 2009, Eitz et al., 2011].

Sketches also arise frequently in shape retrieval applications due to their simplicity and expressiveness. Our focus, the human stick figure sketch, has been used successfully in [Sahillioğlu and Sezgin, 2017] to retrieve 3D human models from large databases. The prominent example in this domain [Eitz et al., 2012] as well as the convolutional neural network based method [Wang et al., 2015] report good performance with gesture drawings when it comes to retrieving humans. These three methods, as well as many other sketch-based retrieval methods [Li et al., 2014], are in general successful on retrieving non-human models as well.

Although human body types under the same pose can be learned easily with moderate amounts of data through statistical shape modeling [Allen et al., 2003, Seo and Magnenat-Thalmann, 2004], this approach requires much greater amounts of

input data to learn plausible shape poses under various deformations [Hasler et al., 2009, Pishchulin et al., 2017]. In addition to the data issue, this family of methods that are based on statistical analysis of human body shape operate directly on vertex positions, which brings the disadvantage that rotated body parts have a completely different representation. This issue is addressed with various heuristics, most successful of which leads to the SMPL model [Loper et al., 2015] that enables 3D human extraction from 2D images [Kanazawa et al., 2018, Omran et al., 2018]. Our learning-based solution requires moderate amount of training data, and also alleviates the rotated part issue by simply populating the input data with 17 other rotated versions of each model.

Chapter 3

OVERVIEW

We have two main objectives: (i) Generating 3D models from a single sketched stick figure, (ii) creating 3D animations between two 3D models, generated from 2D source and target sketches. In addition, we present an application that allows interactive modeling using our algorithm.

Our approach is powered by a Variational Autoencoder network (VAE). We train this network with pairs of 3D and 2D points. The 3D points come from the SCAPE 3D human figure database and TOSCA 3D horse figure database, while the 2D points are obtained by projecting joint positions of these models on a 2D surface. Hence the correspondence information is preserved. Our neural network model ties the 2D and 3D representations through a latent space, which allows us to generate 3D point clouds from 2D stick figures.

The latent space that ties the 2D and 3D representations also acts as a convenient lower dimensional embedding for interpolation and extrapolation. Given a set of target key frames in the form of 2D drawings, we can map them into the lower dimensional embedding space, and then interpolate between them to obtain a number of via points in the embedding space. These via points can then be mapped to 3D through the network to obtain a smooth 3D animation sequence. Furthermore, extrapolation allows extending the animation sequence beyond the target key frames.

Chapter 4

METHODOLOGY

Our method aims to generate static and dynamic 3D models from scratch, that is, we require only the 2D input sketch and no other data such as a rigged 3D model waiting to be reposed. To make this possible, we learn a model that maps 2D stick figures to 3D models.

4.1 Training Data Generation

The original SCAPE [Anguelov et al., 2005] and TOSCA [Bronstein et al., 2008] datasets consist of 72 key 3D meshes of a human actor and 8 key 3D meshes of a horse respectively. They also contains point-to-point correspondence information between these distinct model poses. We use a simple algorithm to extend these datasets by rotating the existing figures with different angles. First, we determine the axes and the angles of the rotation with respect to the original coordinate system shown in Figure 4.1. We ignore the rotation with respect to the x-axis, since stick figures are less likely to be drawn from this view. Next, we rotate the models with respect to the y-axis and z-axis, in a range of -90 degree to 90 at intervals of 30 degrees for the y-axis, and -45 to 45 degrees at intervals of 45 degrees for the z-axis. In the end of this process, we output 21 models per key model in the SCAPE and TOSCA datasets.

Since our network is trained with (2D joints, 3D model) pairs, we also extract 2D joints from a 3D model in a particular perspective. For SCAPE dataset, we designate 11 essential points that alone can describe a 3D human pose. These are the following: forehead(1), elbows(2), hands(2), neck(1), abdomen(1), knees(2) and feet(2). These essential points extends to 12 for TOSCA dataset: forehead(1), shoulder joint(1), hip joint(1) knees(4), feet(4) and tail(1). Since the datasets has

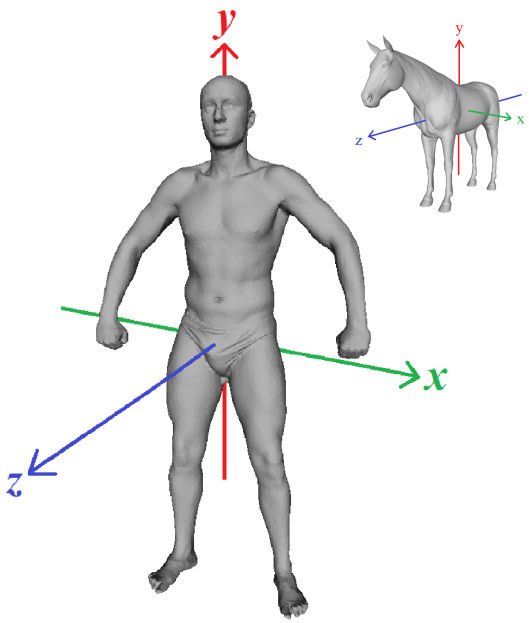

Figure 4.1: Demonstration of sample 3D models from the training datesets with our designated coordinate system.

the point-to-point correspondence information in itself, we select these 3D points in a pilot mesh from each dataset. We project these joints onto a 2D camera plane (x-y in our case) across the entire datasets to create 2D joint projections. In order to be independent from the coordinate system, we represent these points with relative positions (Δx, Δy). We determine two separate specific order in 2D points forming a sketching path with 17 points for our human model and 20 points for our horse model (some joints are visited twice but in reverse direction). Each sketching path determines the order of 2D points that form the input vector of our neural network. The input vector format also handles front/back ambiguity while generating 3D models. We set the first point in the sketching path as the origin, $(0, 0)$ and then we set the remaining points with respect to their relative position to the preceding point.

4.2 Neural Network Architecture

We build upon the work of Girdhar et al. [Girdhar et al., 2016] while designing our neural network architecture. Girdhar et al. aim to learn a vector representation that is *generative*, i.e., it can generate 3D models in the form of voxels; and *predictable*, i.e., it can be predicted by a 2D image. We utilize variational autoencoders rather than standard autoencoders to build our neural network as shown in Figure 4.2. Unlike standard autoencoders, VAEs are generative networks whose latent distribution is regularized during the training in order to be close to a standard Gaussian distribution. This property of VAEs ensures that its latent distribution has a meaningful organization which allows us to generate novel 3D models by sampling in this distribution. In addition to generating novel 3D models, since our framework is capable of learning the vector space around these 3D models, it enables meaningful transitions between them and extrapolations beyond them.

For our training network, we have two encoders and one decoder for each dataset: Encoder3D, Encoder2D, and Decoder. Encoder3D and Decoder together serve as a VAE. Our VAE takes in a 3D point cloud as input, and reconstructs the same model as output. While our VAE learns to reconstruct 3D models, it forces latent distribution of the dataset to approximate normal distribution which makes the latent space interpolatable. Meanwhile, we use our Encoder2D to predict latent vectors of corresponding 3D models from 2D points. In order to provide our latent distribution with similarity information between 3D models, we design this partial architecture for our neural network instead of using a VAE which directly generates 3D models from 2D sketches. Thus, our Encoder3D is capable of learning relations between 3D models rather than 2D sketches while creating a regularized latent distribution. With this method, we aim to explore latent space better and generate more meaningful transitions between 3D models.

4.2.1 Encoder3D-Decoder VAE Network Architecture

Our VAE takes 12500×3 point cloud of human 3D model or 19248×3 point cloud of horse 3D model as input. Encoder3D contains two fully connected layers and its

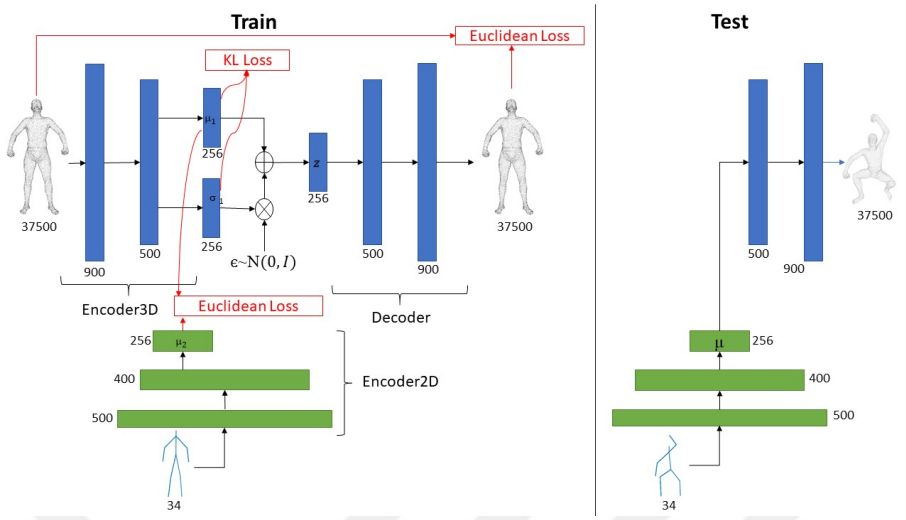

Figure 4.2: Our neural network architecture. **(a) Train Network:** We train this network with (3D point cloud, 2D points of stick figure) pairs during training time. It consists of a VAE: Encoder3D and Decoder consecutively, and another external encoder: Encoder2D. We use regression loss from the output of Encoder2D to the mean vector of the VAE in addition to standard losses of VAE. **(b) Test Network:** We remove Encoder3D and reparameterization layer from our VAE and use Encoder2D-Decoder as our network in our experiments.

outputs are a mean vector and a deviation vector. We use ReLU as an activation layer in all the internal layers. There is no activation layer in the output layers. Our Decoder takes the latent vector z as input. It also consists of two fully connected layers with a ReLU activation layer and one fully connected output layer with a *tanh* activation layer. It gives a reconstructed point cloud of the input 3D model as output.

We train our VAE with the standard KL divergence and reconstruction losses. The total loss for our VAE is given in Equation 4.1.

$$L_{\text{VAE}} = \alpha \frac{1}{BN} \sum_{i=1}^{B} \sum_{j=1}^{N} (x_{ij} - \hat{x}_{ij})^2 + D_{\text{KL}}(q(z|x)||p(z)) \tag{4.1}$$

In Equation 4.1, α is a parameter to balance between the reconstruction loss and KL divergence loss, B is the training batch size, N is the 1D dimension of vectorized point cloud (37500 in our case for human models), x_{ij} is the j-th dimension of i-th model in the training data batch, $\hat{x}_{ij}$ is the j-th dimension of model i's output from

our VAE, z is the reparameterized latent vector, $p(z)$ is the prior probability, $q(z|f)$ is the posterior probability, and D_{KL} is KL divergence.

4.2.2 Mapping 2D Sketch Points to Latent Vector Space

Our Encoder2D learns to predict the latent vectors of 3D models that corresponds to 2D sketches as discussed. It takes 17×2 points of human stick figure or 20×2 points of horse stick figure as input to map it into mean vector. It has the same structure with Encoder3D except its input and internal fully connected layers' dimensions. In the test case we use Encoder2D and Decoder as a standard autoencoder. Decoder takes the mean vector output of Encoder2D as its input and generates 3D point cloud as the output.

We train our Encoder2D with mean square loss to regress 256D representation of mean vector given by pre-trained Encoder3D. The loss for our Encoder2D is given in Equation 4.2.

$$L_{\mathrm{2D}} = \frac{1}{BZ} \sum_{i=1}^{B} \sum_{j=1}^{Z} (\mu_i^1 - \mu_i^2)^2 \tag{4.2}$$

In Equation 4.2, B is the training batch size, Z is the dimension of latent space (256 in our case), μ_{ij}^1 is the j-th dimension of mean vector produced by Encoder3D to i-th model in training batch, and μ_{ij}^2 is the j-th dimension of mean vector produced by Encoder2D to i-th model in the training batch.

4.3 Training Details

We follow a three-stage process to train our network for each dataset. (i) We train our variational autoencoder independently with the loss function in Equation (4.1). We run this stage for 300 epochs. (ii) We train our Encoder2D with the loss function in Equation (4.2) using Encoder3D trained from (i). Specifically, we train our Encoder2D to regress the latent vector produced from the pre-trained Encoder3D for the input 3D model. We run this stage for 300 epochs. (iii) We use both losses

jointly to finetune our framework. We run this stage for 200 epochs. It takes about two days to complete whole training session.

For the experiments throughout this paper, we set $\alpha = 10^5$. We set the prior probability over latent variables as a standard normal distribution, $p(z) = \mathcal{N}(z; 0, I)$. We set the learning rate as 10^{-3} and 10^4 for our human model generation and horse model generation networks respectively. We use the Adam Optimizer as our optimizer in training.

4.4 User Interface

As we have explained in prior sections, our method can be used in a variety of applications in different fields such as character generation and making quick animations. In order to better utilize these applications we propose a user interface with facilitative properties that enables users to perform our method in a better manner (Figure 4.3). Our user interface acts as an agent that ensures the input-output communication between the user and our neural network. The user first choose whether to generate human 3D models or horse 3D models. Then, the user can choose whether to generate 3D models from 2D stick figure sketches or create animations between the source and target stick figure sketches. While it takes about one second to process a sketch input for generation of the corresponding 3D model, this time extends approximately to five seconds for producing animation between (interpolation) and beyond (extrapolation) sketch inputs.

Our user interface takes a sketch input from the user via an embedded 2D canvas (right part). The collected sketch is transformed into a map of the joint locations as an input to our neural network by requiring users to sketch in a predetermined path. The user interface then shows the 3D model output produced by the neural network on the embedded 3D canvas (left part). Although our output is a 3D point cloud, for better visualization, our user interface utilizes the mesh information that already exists in the SCAPE and TOSCA datasets. Generated point clouds are consequently combined with this information to display 3D models as surfaces with appropriate rendering mechanisms such as shading and lighting.

Since we trained an abundance of neural networks until achieving the best one with the optimal parameters, our user interface showed two different 3D model outputs coming from two different neural networks for comparisons during the development phase. The interface in Figure 4.3 belongs to our release mode where only the promoted 3D output is displayed.

We construct our user interface such that it is purified from unnatural interaction tools such as buttons and menus. Generation process starts as soon as the last stroke is drawn without forcing the user to hit the start button. We provide brief information in text that describes the canvas organization. To make the interaction more fluent, we add a simple "scribble" detector to understand the undo operation.

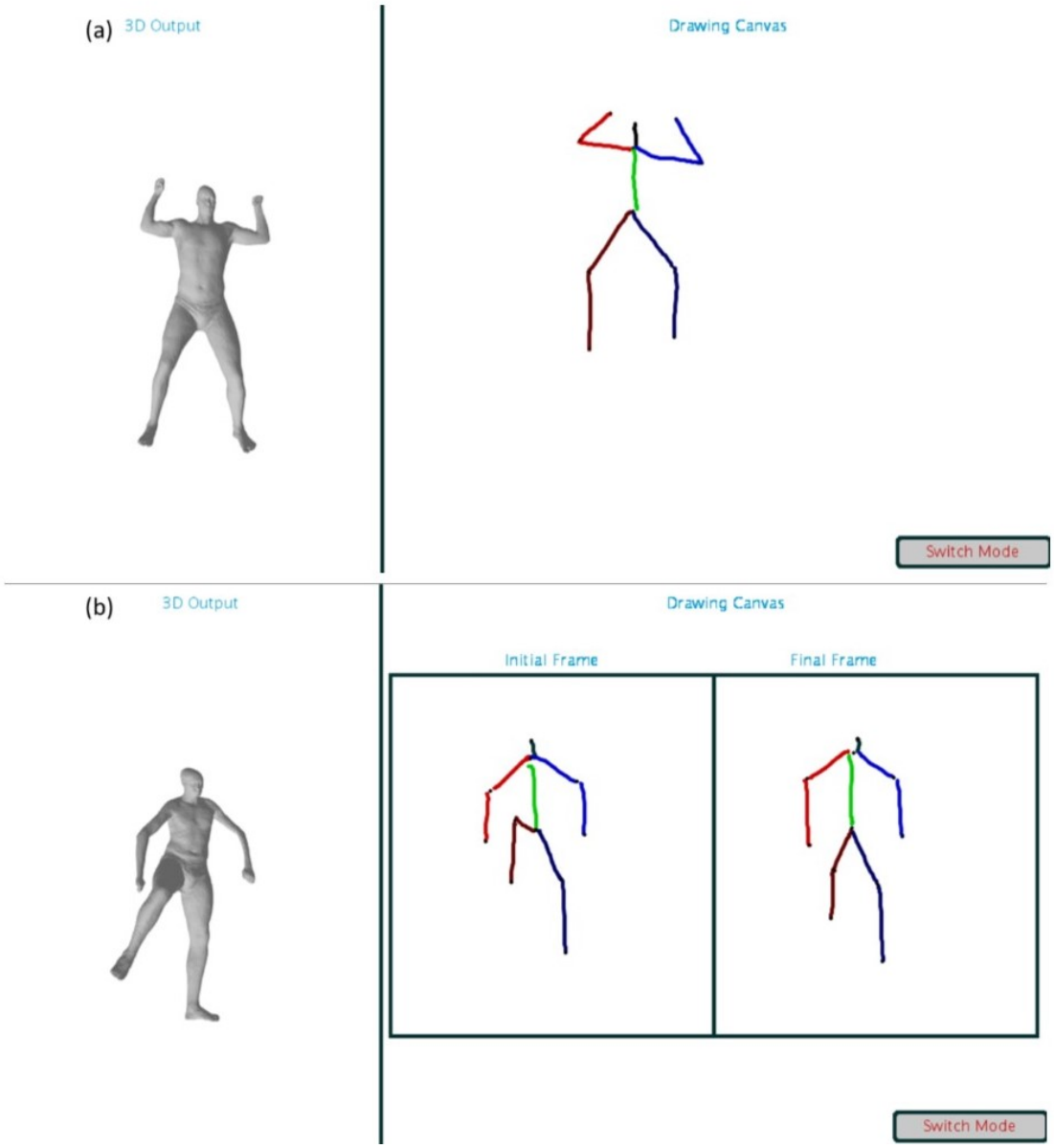

Figure 4.3: Screenshots from our user interface. (a) 3D model generation mode. (b) 3D animation generation mode.

Chapter 5

EXPERIMENTS AND RESULTS

In this section, we first evaluate our framework qualitatively and quantitatively. We evaluate three tasks performed by our framework. (i) Generating 3D models from 2D stick figure sketches. (ii) Generating 3D animations between source and target stick figure sketches. (iii) Performing simple extrapolation beyond stick figures.

5.1 Framework Evaluation

5.1.1 Standard Autoencoder Baseline

To quantitatively justify our preference on variational autoencoders, we design a standard autoencoder (AE) baseline with similar dimensions and activation functions (Figure 5.1). We train this network on SCAPE dataset for 300 epochs with Euclidean loss between generated 3D models and ground truth ones. We compare the per-vertex reconstruction loss on validation set consisting held-out 3D models with our VAE network and standard AE network trained with human models. The results on Table 5.1 shows that our VAE network outperforms AE network, exploiting latent space more efficiently to enhance the generation quality of novel 3D models.

Table 5.1: Per-vertex reconstruction error on validation data with different network architectures. VAE represents our method.

Method	Z (Latent Dim.)	Mean (x10^{-3})	Std (x10^{-3})
AE	256	2.996	1.568
VAE	256	2.118	2.598

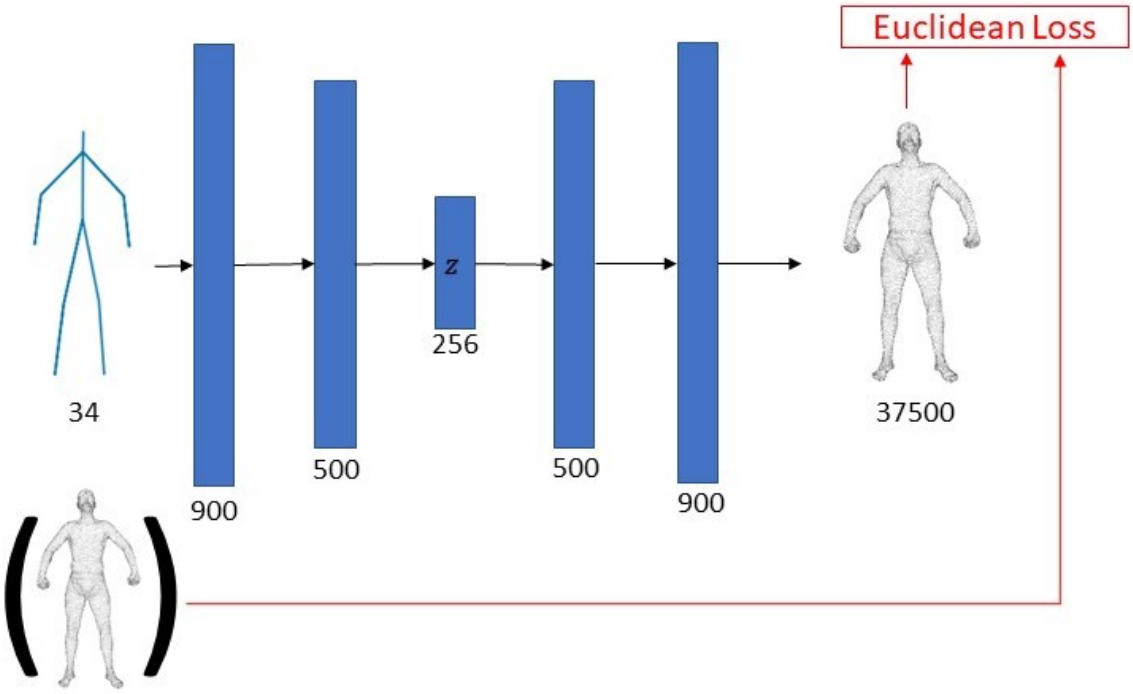

Figure 5.1: Neural network architecture for standard autoencoder baseline.

5.1.2 Latent Dimensions

We evaluate different latent dimensions for training our VAE framework with SCAPE dataset using per-vertex reconstruction loss on validation set. The results on Table 5.2 show that using 256 dimensions improves generation quality compared to lower dimensions. Higher dimensions lead to overfitting. We use 256 dimensions for the following experiments.

Table 5.2: Per-vertex reconstruction error on validation data with different latent dimensions. 256 is our promoted latent space dimension.

Method	Z (Latent Dim.)	Mean (x10^{-3})	Std (x10^{-3})
VAE	128	3.887	3.802
VAE	256	2.118	2.598
VAE	512	10.830	6.888

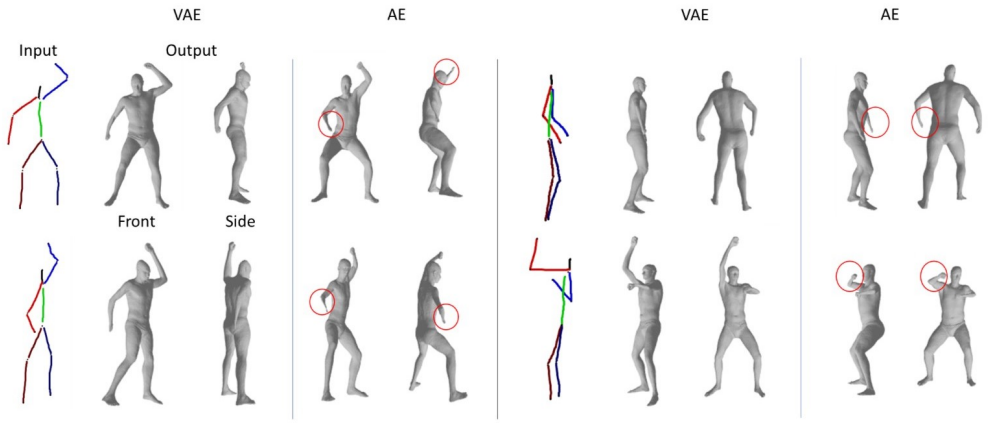

Figure 5.2: Generated 3D human models in front and side views for given sketches using our VAE network and standard AE network. Flaws are highlighted with red circles.

5.2 Generating 3D models from 2D stick figure sketches

To evaluate the generation ability, we feed 2D human stick figure sketches as input to our framework. Our user interface is used for the sketching activity. In Figure 5.2, we compare generation results for sample sketches with our VAE network and standard AE baseline. These results show that standard AE network generates models with anatomical flaws (collapsed arm in Figure 5.2) and deficient body orientations. Our VAE network produces compatible models of high quality.

We compare the generation ability of our framework with a recent study [Kanazawa et al., 2018] that predicts 3D shape from 2D joints of human. While our framework outputs 3D point clouds, the method described in [Kanazawa et al., 2018], tries to fit a statistical body shape model, SMPL [Bogo et al., 2016], on 2D joints. Their learned statistical model is a male who is in a different shape than our male model as shown in the second column of Figure 5.3. We take the visuals reported in their paper and draw the corresponding human stick figures for a fair comparison. Despite being restricted to the reported poses in [Kanazawa et al., 2018], our method compares favorably. The sitting pose, for instance, which is not quite captured by their statistical model shows in an inaccurate 3D sitting while our 3D model sits successfully (Figure 5.3 - top row).

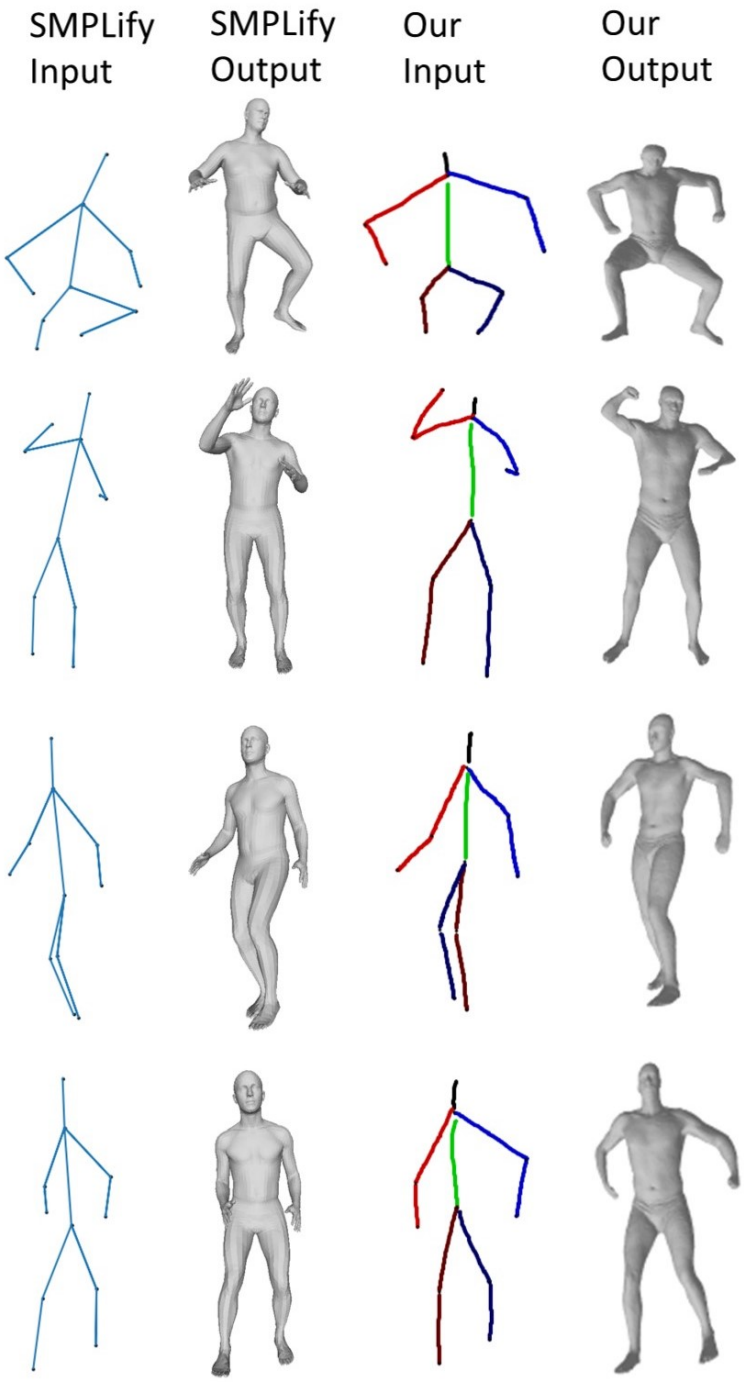

Figure 5.3: Qualitative comparison of our method (columns 3 and 4) with [Kanazawa et al., 2018] (columns 1 and 2).

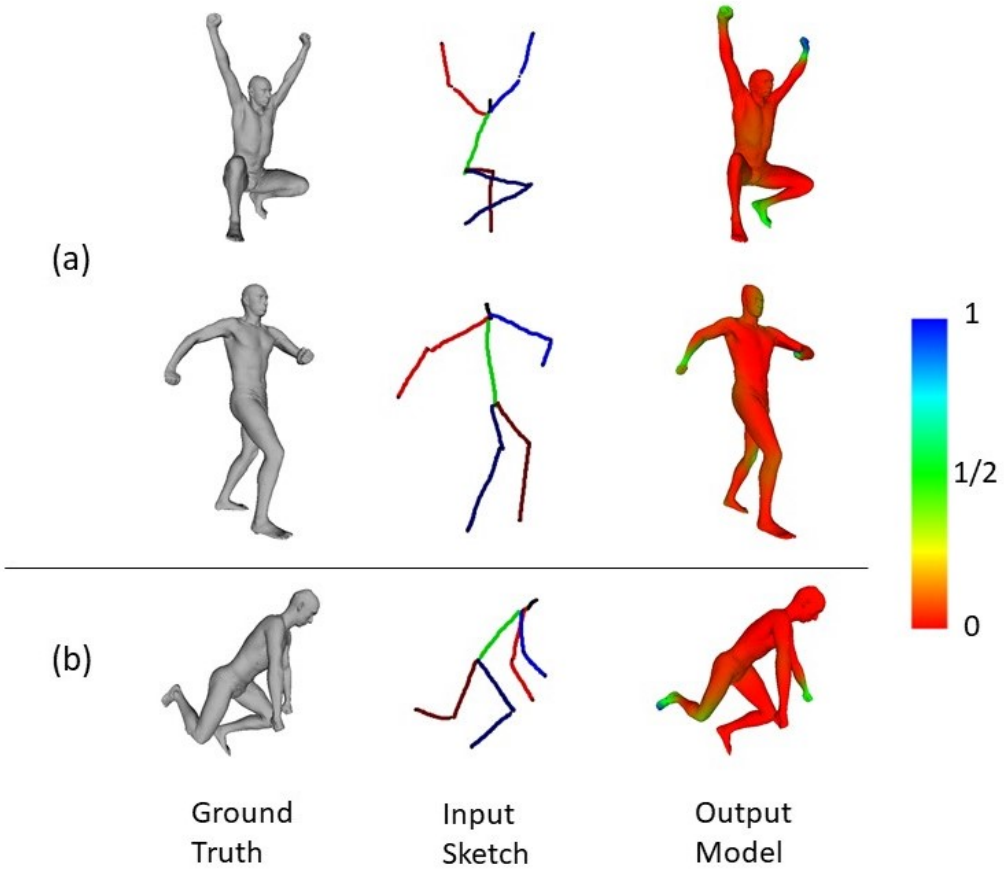

Figure 5.4: Generation results for (a) two human stick figure sketches observed in training and (b) an unobserved human stick figure sketch. The generated models (last column) are colored with respect to the distance map to the ground truth (first column). Distance values are normalized between 0 and 1.

We also compare our 3D human model generation ability to the ground-truth by feeding sketches that resemble 3 of the 72 SCAPE poses used during training. Two of these poses were observed during training at the same orientations as we draw them in the first two rows of Figure 5.4, yet the last one is drawn with a new orientation that was not observed during training (Figure 5.4-last row). Consequently, we had to align the generated model with the ground-truth model via ICP alignment [Besl and McKay, 1992] for the last row prior to the comparison. We observe almost perfect replications of the SCAPE poses in all cases as depicted by the colored difference maps.

In order to show generalization ability of our method with different 3D model

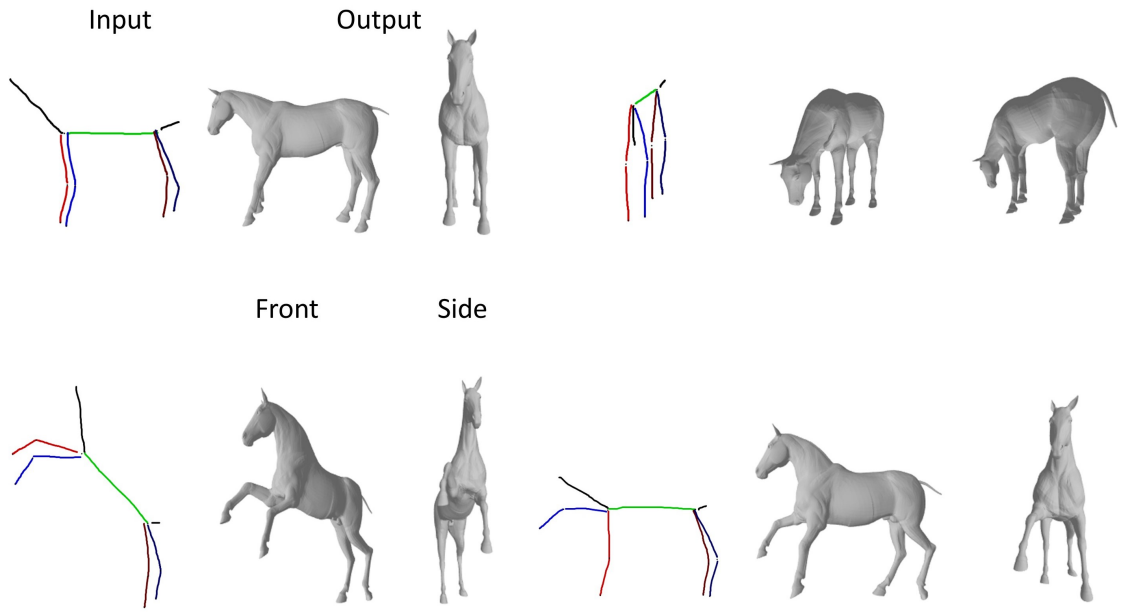

Figure 5.5: Generated 3D horse models in front and side views for given sketches using our VAE network.

types, we evaluate our VAE framework which is trained on the TOSCA dataset. To evaluate the generation ability with 3D horse models, we feed 2D horse stick figure sketches as input to our framework. In Figure 5.5, we show generation results for sample sketches with our VAE network.

5.3 Generation of Dynamic 3D Models

5.3.1 Interpolation

We test the capability of our network to generate 3D human animations between source and target stick figures. To accomplish this, our framework takes two stick figure sketches via our user interface. The framework then encodes the stick figures into the embedding space to extract their latent variables. We create a list of latent variables by linearly interpolating between the source and the target latent variables. We feed this list as an input to our decoder, with each element of the list being fed one by one to produce the desired 3D model sequence. In Figure 5.6 (a), we compare our results with direct linear interpolation between source and target 3D model output.

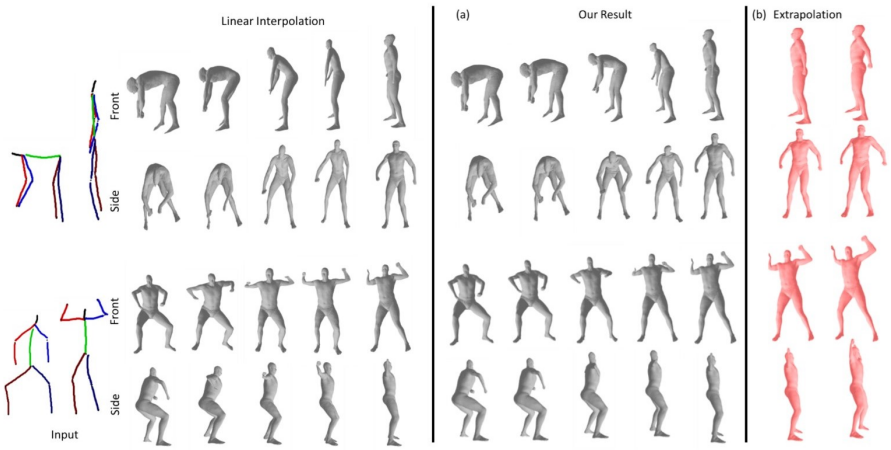

Figure 5.6: Qualitative comparison with linear interpolation. (a) Produced 3D model sequences for given sketches using our network are better than linear interpolation results. (b) Extrapolation results for 3D model sequences are given as red models.

Our results show that the interpolation in embedding space can avoid anatomical errors which usually occur in methods using direct linear interpolation. Further interpolation results can be found in the teaser image and the accompanying video.

We also compare our 3D human model generation ability with interpolation at the sketch level. In order to implement this, we design an algorithm that interpolates between source and target stick figure sketches. Our algorithm first finds the best rotation vector between each component of source and target stick figure sketches using the proposed method in [Arun et al., 1987]. Then it generates interpolated rotations in the form of quaternion using quaternion spherical interpolation. We obtain the list of interpolated sketches with this interpolated rotations. We feed this sketch list as an input to our decoder, with each element of the list being fed one by one to produce the desired 3D model sequence. In Figure 5.7, we compare our embedding space interpolation results with sketch level interpolation. The 3D model generation results on interpolated sketches shows the high adaptation capacity of our framework to novel sketches. However, interpolation in embedding space results in more meaningful transition between source and target poses. For example in sketch interpolation results in Figure 5.7, the leg of middle 3D model unnecessarily

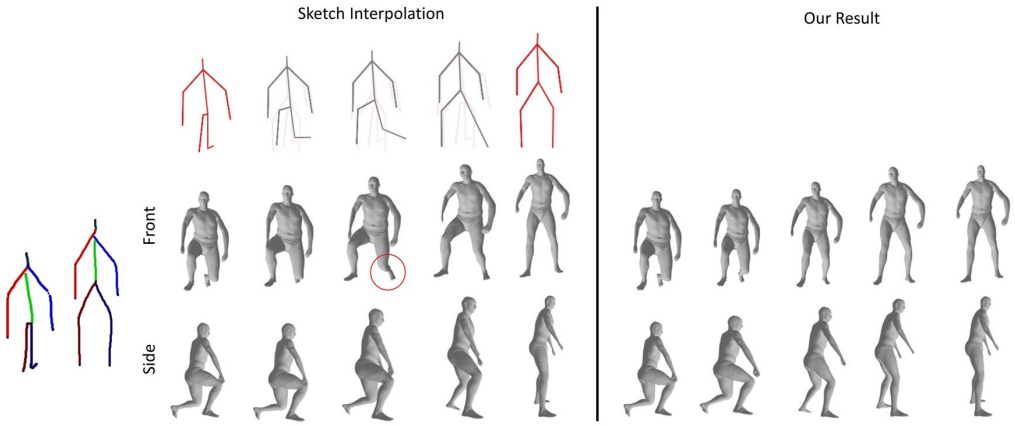

Figure 5.7: Qualitative comparison with interpolation at the sketch level. The transition between produced 3D model sequences using interpolation in embedding space are more meaningful than interpolation at sketch level.

turns right due to similar pose information coming from the input sketch as pointed out in red circle.

5.3.2 Extrapolation

Our results show that the interpolation vector in the embedding space is capable of reasonably representing the motion in real world actions. To improve upon this idea, we exploit the learned interpolation vector in order to predict future movements. We show our results for extrapolation in Figure 5.6 (b). This figure shows that the learned interpolation vector between two 3D shapes contains meaningful information of movement in 3D. Further extrapolation results can be found in the teaser image and the accompanying video.

5.4 Timing

We finish our evaluations with our timing information. The closest work to our framework reports 10 minutes of production time for an amateur user to create 3D faces from 2D sketches [Han et al., 2017]. In our system, on the other hand, it takes an amateur user 10 seconds of stick figure drawing and 1 second of algorithm processing to create 3D bodies from 2D sketches. There are two main reasons of

this remarkable performance advantage of our framework: i) Human bodies can be naturally represented with easy-to-draw stick figures whereas faces cannot. The simplicity and expressiveness make the learning easier and more efficient. ii) Our deep regression network is significantly less complex than the one employed in [Han et al., 2017].

Our application runs on a PC with 8GB ram and i7 2.80GHz CPU. Training of our model is done on Tesla K40m GPU and took about 2 days (800 epochs total).

Chapter 6

LIMITATIONS

The proposed system has several limitations that are listed as follows:

- Training of our network is dependent on the existing 3D shapes in the dataset. Our network cannot learn vastly different shapes than existing ones: it produces incompatible 3D models with sketch inputs that are not closely represented in the dataset. For example, our network does not correctly capture the arm orientation for the right model in Figure 6.1.

- The system can only generate human shapes because of the content of the dataset.

- The system can produce articulated shapes. Although it can twist and bend human body in a reasonable way, it can not, for instance, stretch or resize a body part.

- The system benefits from the one-to-one correspondence information of the dataset. Thus, the quality of results depends on this information.

- Since our network takes its input in a specific order, our user interface constrains users to sketch in that order. Users cannot sketch stick figures in an arbitrary order.

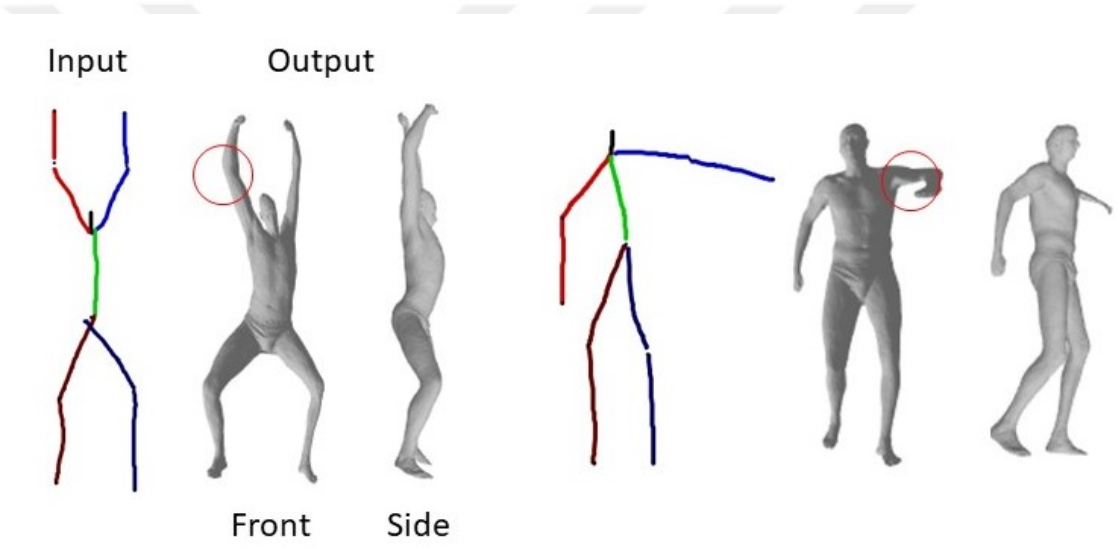

Figure 6.1: Failure cases of our framework. Our framework generates anatomically unnatural results if the input sketch has disproportionate body parts (left), or it is significantly different than the ones used during training (right).

Chapter 7

CONCLUSIONS

In this paper, we presented a deep learning based framework that is capable of generating 3D models from 2D stick figure sketches and producing dynamic 3D models between (interpolation) and beyond (extrapolation) two given stick figure sketches. Unlike existing methods, our method requires neither a statistical body shape model nor a rigged 3D character model. We demonstrated that our framework not only gives reasonable results on generation, but also compares favorably with existing approaches. We further supported our framework with a well-designed user interface to make it practical for a variety of applications.

Chapter 8

FUTURE WORK

Potential future work directions that align with our proposed system are described as follows. Human stick figures used in this paper can be generalized to any other shape using their skeletons. Consequently, an automated skeleton extraction algorithm would enable further training of our network, which in turn extends our solution to non-human objects. Voxelization of our input data would spare us from the one-to-one correspondence information requirement, which in turn would enable our interpolation scheme to morph from different object classes that do not often have this type of information, e.g., from cat to giraffe. Automatic one-to-one correspondence computation [Kim et al., 2011, Sahillioğlu and Yemez, 2013, Sahillioğlu, 2018] can also be considered to avoid voxelization. Latent space can be exploited in a better manner in order to obtain a more sophisticated extrapolation algorithm than the basic one we introduced in this paper. New sketching cues can be designed and incorporated into our network to be able to produce body types different than the one used during training, e.g., training with the fit SCAPE actor and production with a obese actress.

BIBLIOGRAPHY

[Allen et al., 2003] Allen, B., Curless, B., and Popović, Z. (2003). The space of human body shapes: reconstruction and parameterization from range scans. *ACM transactions on graphics (TOG)*, 22(3):587–594.

[Anastacio et al., 2006] Anastacio, F., Sousa, M. C., Samavati, F., and Jorge, J. A. (2006). Modeling plant structures using concept sketches. In *Proceedings of the 4th international symposium on Non-photorealistic animation and rendering*, pages 105–113. ACM.

[Anguelov et al., 2005] Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., and Davis, J. (2005). Scape: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416.

[Arun et al., 1987] Arun, K. S., Huang, T. S., and Blostein, S. D. (1987). Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):698–700.

[Besl and McKay, 1992] Besl, P. and McKay, N. (1992). Method for registration of 3-d shapes. *Sensor Fusion IV: Control Paradigms and Data Structures*, 1611.

[Bessmeltsev et al., 2016] Bessmeltsev, M., Vining, N., and Sheffer, A. (2016). Gesture3d: posing 3d characters via gesture drawings. *ACM Transactions on Graphics (TOG)*, 35(6):165.

[Bogo et al., 2016] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., and Black, M. J. (2016). Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing.

[Bronstein et al., 2008] Bronstein, A., Bronstein, M., and Kimmel, R. (2008). *Numerical Geometry of Non-Rigid Shapes*. Springer Publishing Company, Incorporated, 1 edition.

[Chen et al., 2009] Chen, T., Cheng, M.-M., Tan, P., Shamir, A., and Hu, S.-M. (2009). Sketch2photo: Internet image montage. *ACM Transactions on Graphics (TOG)*, 28(5):124.

[Choi et al., 2012] Choi, M. G., Yang, K., Igarashi, T., Mitani, J., and Lee, J. (2012). Retrieval and visualization of human motion data via stick figures. In *Computer Graphics Forum*, volume 31, pages 2057–2065. Wiley Online Library.

[Davis et al., 2003] Davis, J., Agrawala, M., Chuang, E., Popović, Z., and Salesin, D. (2003). A sketching interface for articulated figure animation. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 320–328. Eurographics Association.

[Delanoy et al., 2018] Delanoy, J., Aubry, M., Isola, P., Efros, A., and Bousseau, A. (2018). 3d sketching using multi-view deep volumetric prediction. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 1(21).

[Eitz et al., 2012] Eitz, M., Richter, R., Boubekeur, T., Hildebrand, K., and Alexa, M. (2012). Sketch-based shape retrieval. *ACM Trans. Graph.*, 31(4):31–1.

[Eitz et al., 2011] Eitz, M., Richter, R., Hildebrand, K., Boubekeur, T., and Alexa, M. (2011). Photosketcher: interactive sketch-based image synthesis. *IEEE Computer Graphics and Applications*, 31(6):56–66.

[Fu et al., 2007] Fu, H., Wei, Y., Tai, C.-L., and Quan, L. (2007). Sketching hairstyles. In *Proceedings of the 4th Eurographics workshop on Sketch-based interfaces and modeling*, pages 31–36. ACM.

[Girdhar et al., 2016] Girdhar, R., Fouhey, D. F., Rodriguez, M., and Gupta, A.

(2016). Learning a predictable and generative vector representation for objects. *CoRR*, abs/1603.08637.

[Guay et al., 2013] Guay, M., Cani, M.-P., and Ronfard, R. (2013). The line of action: an intuitive interface for expressive character posing. *ACM Transactions on Graphics (TOG)*, 32(6):205.

[Hahn et al., 2015] Hahn, F., Mutzel, F., Coros, S., Thomaszewski, B., Nitti, M., Gross, M., and Sumner, R. W. (2015). Sketch abstractions for character posing. In *Proceedings of the 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 185–191. ACM.

[Han et al., 2017] Han, X., Gao, C., and Yu, Y. (2017). Deepsketch2face: A deep learning based sketching system for 3d face and caricature modeling. *ACM Transactions on Graphics*, 36(4):126.

[Hasler et al., 2009] Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., and Seidel, H.-P. (2009). A statistical model of human pose and body shape. *Computer Graphics Forum*, 28(2):337–346.

[Igarashi et al., 1999] Igarashi, T., Matsuoka, S., and Tanaka, H. (1999). Teddy: a sketching interface for 3d freeform design. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 409–416. ACM Press/Addison-Wesley Publishing Co.

[Kanazawa et al., 2018] Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2018). End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Karpenko and Hughes, 2006] Karpenko, O. A. and Hughes, J. F. (2006). Smoothsketch: 3d free-form shapes from complex sketches. *ACM Transactions on Graphics (TOG)*, 25(3):589–598.

[Kim et al., 2011] Kim, V. G., Lipman, Y., and Funkhouser, T. (2011). Blended intrinsic maps. *ACM Transactions on Graphics (TOG)*, 30(4):79.

[Lee and Funkhouser, 2008] Lee, J. and Funkhouser, T. A. (2008). Sketch-based search and composition of 3d models. In *SBM*, pages 97–104.

[Li et al., 2014] Li, B., Lu, Y., Li, C., Godil, A., Schreck, T., Aono, M., Burtscher, M., Fu, H., Furuya, T., Johan, H., et al. (2014). Shrec'14 track: Extended large scale sketch-based 3d shape retrieval. In *Eurographics workshop on 3D object retrieval*, volume 2014. .

[Li et al., 2017] Li, C., Pan, H., Liu, Y., Tong, X., Sheffer, A., and Wang, W. (2017). Bendsketch: modeling freeform surfaces through 2d sketching. *ACM Transactions on Graphics*, 36(4):125.

[Lin et al., 2012] Lin, J., Igarashi, T., Mitani, J., Liao, M., and He, Y. (2012). A sketching interface for sitting pose design in the virtual environment. *IEEE transactions on visualization and computer graphics*, 18(11):1979–1991.

[Loper et al., 2015] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248.

[Miao et al., 2015] Miao, Y., Hu, F., Zhang, X., Chen, J., and Pajarola, R. (2015). Symmsketch: Creating symmetric 3d free-form shapes from 2d sketches. *Computational Visual Media*, 1(1):3–16.

[Nealen et al., 2007] Nealen, A., Igarashi, T., Sorkine, O., and Alexa, M. (2007). Fibermesh: designing freeform surfaces with 3d curves. *ACM transactions on graphics (TOG)*, 26(3):41.

[Olsen et al., 2009] Olsen, L., Samavati, F. F., Sousa, M. C., and Jorge, J. A. (2009). Sketch-based modeling: A survey. *Computers & Graphics*, 33(1):85–103.

[Omran et al., 2018] Omran, M., Lassner, C., Pons-Moll, G., Gehler, P. V., and Schiele, B. (2018). Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision, 3DV 2018, Verona, Italy, September 5-8, 2018*, pages 484–494.

[Öztireli et al., 2013] Öztireli, A. C., Baran, I., Popa, T., Dalstein, B., Sumner, R. W., and Gross, M. (2013). Differential blending for expressive sketch-based posing. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 155–164. ACM.

[Pishchulin et al., 2017] Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., and Schiele, B. (2017). Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67:276–286.

[Sahillioğlu, 2018] Sahillioğlu, Y. (2018). A genetic isometric shape correspondence algorithm with adaptive sampling. *ACM Transactions on Graphics (TOG)*, 37(5):175.

[Sahillioğlu and Sezgin, 2017] Sahillioğlu, Y. and Sezgin, M. (2017). Sketch-based articulated 3d shape retrieval. *IEEE computer graphics and applications*, 37(6):88–101.

[Sahillioğlu and Yemez, 2013] Sahillioğlu, Y. and Yemez, Y. (2013). Coarse-to-fine isometric shape correspondence by tracking symmetric flips. *Computer Graphics Forum*, 32(1):177–189.

[Seki and Igarashi, 2017] Seki, S. and Igarashi, T. (2017). Sketch-based 3d hair posing by contour drawings. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 29. ACM.

[Seo and Magnenat-Thalmann, 2004] Seo, H. and Magnenat-Thalmann, N. (2004). An example-based approach to human body manipulation. *Graphical Models*, 66(1):1–23.

[Sezgin et al., 2006] Sezgin, T. M., Stahovich, T., and Davis, R. (2006). Sketch based interfaces: early processing for sketch understanding. page 22.

[Shin and Igarashi, 2007] Shin, H. and Igarashi, T. (2007). Magic canvas: interactive design of a 3-d scene prototype from freehand sketches. In *Proceedings of Graphics Interface 2007*, pages 63–70. ACM.

[Shtof et al., 2013] Shtof, A., Agathos, A., Gingold, Y., Shamir, A., and Cohen-Or, D. (2013). Geosemantic snapping for sketch-based modeling. *Computer graphics forum*, 32(2pt2):245–253.

[Smirnov et al., 2019] Smirnov, D., Bessmeltsev, M., and Solomon, J. (2019). Deep sketch-based modeling of man-made shapes. *ArXiv*, abs/1906.12337.

[Tsang et al., 2004] Tsang, S., Balakrishnan, R., Singh, K., and Ranjan, A. (2004). A suggestive interface for image guided 3d sketching. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 591–598. ACM.

[Wang et al., 2015] Wang, F., Kang, L., and Li, Y. (2015). Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1883.

[Wei et al., 2011] Wei, X. K., Chai, J., et al. (2011). Intuitive interactive human-character posing with millions of example poses. *IEEE Computer Graphics and Applications*, 31(4):78–88.

[Xu et al., 2013] Xu, K., Chen, K., Fu, H., Sun, W.-L., and Hu, S.-M. (2013). Sketch2scene: sketch-based co-retrieval and co-placement of 3d models. *ACM Transactions on Graphics (TOG)*, 32(4):123.

[Yang et al., 2005] Yang, C., Sharon, D., and van de Panne, M. (2005). Sketch-based modeling of parameterized objects. In *SIGGRAPH Sketches*, page 89. Citeseer.

[Zeleznik et al., 1996] Zeleznik, R., Herndon, K., and Hughes, J. (1996). Sketch: an interface for sketching 3d scenes. In *Proceedings of the SIGGRAPH*.

Appendix A

Appendix goes here.