BézierSketch: A generative model for scalable vector sketches

Ayan Das[1,2], Yongxin Yang[1,2], Timothy Hospedales[1,3], Tao Xiang[1,2], and Yi-Zhe Song[1,2]

[1] SketchX, CVSSP, University of Surrey, United Kingdom
{a.das,yongxin.yang,t.xiang,y.song}@surrey.ac.uk
[2] iFlyTek-Surrey Joint Research Centre on Artificial Intelligence
[3] University of Edinburgh, United Kingdom
t.hospedales@ed.ac.uk

Abstract. The study of neural generative models of human sketches is a fascinating contemporary modeling problem due to the links between sketch image generation and the human drawing process. The landmark SketchRNN provided breakthrough by sequentially generating sketches as a sequence of waypoints. However this leads to low-resolution image generation, and failure to model long sketches. In this paper we present BézierSketch, a novel generative model for fully *vector* sketches that are automatically scalable and high-resolution. To this end, we first introduce a novel inverse graphics approach to stroke embedding that trains an encoder to embed each stroke to its best fit Bézier curve. This enables us to treat sketches as short sequences of paramaterized strokes and thus train a recurrent sketch generator with greater capacity for longer sketches, while producing scalable high-resolution results. We report qualitative and quantitative results on the *Quick, Draw!* benchmark.

Keywords: Sketch generation, Scalable graphics, Bézier curve

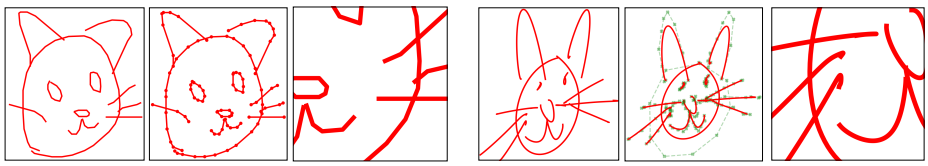


Fig. 1: Left: SketchRNN [8] generates sketches by sampling waypoints (red dots) which lead to coarse images upon zoom. Right: Our BézierSketch samples smooth curves (green control points) thus providing scalable *vector* graphic generation.

1 Introduction

Generative neural modeling of images [6, 12] is now an established research area in contemporary machine learning and computer vision. Rapid progress has been made in generating photos [11, 24], with effort being focused on fidelity, diversity,

and resolution of image generation, along with stability of training; as well as sequential models for text and video [2, 31]. Generative modeling of human *sketches* in particular has recently gained interest, along with other applications of sketch analysis such as recognition [34, 33], retrieval [28, 21, 4] and forensics [13] – all facilitated by the growth of large scale sketch datasets [8, 28].

Sketch generation provides an excellent opportunity to study sequential generative models, and is particularly fascinating due to the potential to establish links between learned generative models and human sketching – a communication modality that comes innately to children, and has existed for millennia. Recent breakthroughs in this area include SketchRNN [8], which provided the first neural generative sequential model for sketch images, and Learn2Sketch [30] which provided the first conditional image to sequential sketch model. While conventional image generation models focus on producing ever-larger pixel arrays in high fidelity, these methods aim to model sketches using a more human-like representation consisting of a collection of strokes.

SketchRNN [8], the landmark neural sketch generation algorithm, treats sketches as a digitized sequence of 2D points on a drawing canvas sampled along the trajectory of the ink-flow. This model of sketches has several issues, however: It is inefficient, due to the dense representation of redundant information like highly correlated temporal samples; and as sketches are ultimately pixels on a grid, it is prone to sampling noise. Crucially it provides limited graphical scalability: SketchRNN sets out to achieve vector graphic generation (and claims to achieve this). However it does not generate truly scalable vector graphs as required by applications such as digital art. Since generated sketches are composed of dense line segments, its samples are only somewhat smoother than raster graphics (Fig. 1). Finally, it suffers from limited capacity. Because it models sketches as a sequence of pixels, it is limited in the length of sketch it can model before the underlying recurrent neural network begins to run out of capacity.

In this paper we propose a fundamental paradigm change in the representation of sketches that enables the above issues to be addressed. Specifically, we aim to represent sketches in terms of parameterized smooth curves [27]. These provide a scalable representation of a finite length curve using few *Control Points*. From a large family of parametric curves, we choose Bézier curves due to their simple structure. In order to train a generative model of human sketches with this representation, the key question is how to encode human sketches as parameterized curves. To this end, a key technical contribution is a vision-as-inverse-graphics [14, 26, 5] approach, that learns to embed human sketch strokes as interpretable parameterized Bézier curves. We train BézierEncoder in an inverse-graphics manner by learning to reconstruct strokes through a white-box graphics (Bézier) decoder. Given this new low-dimensional stroke representation, we then train BézierSketch to generate sketches. Our stroke-level generative model requires many fewer iterations than the segment-level SketchRNN, and thus provides better generation of longer sketches, while providing high-resolution scalable vector-graphic sketch generation (Fig. 1).

In summary, the contributions of our work are: (1) BézierEncoder, a novel inverse-graphics approach for mapping strokes to parameterized Béziers, (2) BézierSketch, a sequential generative model for sketches that produces high-resolution and low-noise vector graphic samples with improved scalability to longer sketches compared to the previous state of the art SketchRNN.

2 Related Work

Parameterized Curves Bézier curves are a powerful tool in the field of computer graphics and are extensively used in interactive curve and surface design [27], as are a more general family of curves known as *Splines* [3]. Optimization algorithms to fit Bézier curves and Splines from data have been studied. Few specially crafted algorithms do exist specifically for cubic Bézier curves [29, 20]. However the challenge for most curve and spline-fitting methods is the existence of latent variables t that correspond training points and the location of their projection onto the curve. This leads to two-stage alternating algorithms for separately optimizing the curve parameters (control points) and latent parameter t [17, 22]. Importantly, such methods [17, 22] including few promising ones [35] require expensive *per-sample* alternating optimization, or iterative inference in expensive generative models [25, 15] which make them unsuitable for large scale or online applications. In contrast, we uniquely take the approach of learning a neural network that maps strokes to Bézier curves in a single shot. This neural encoder is a model that needs to be trained, but unlike per-sample optimization approaches, it is inductive. So once trained it can provide one-shot estimation of curve parameters and point association from an input stroke.

Generative Models Generative models have been studied extensively in the machine learning literature, often in terms of density estimation with directed [23, 1] or undirected [10] graphical models. Research in this field accelerated after the emergence of Generative adversarial networks (GAN) [6], Variational Autoencoder (VAE) [12] and their derivatives. Handling sequences are of particular importance and hence specialized algorithms [2, 31] were developed. Although RNNs have been successfully used for generating handwriting [7] without variational training, these methods lacked flexibility in terms of generation quality. The emergence of VAE and variational training methods allows the fusion of RNNs with variational objective led to the first successful generative sequence model [2] in the domain of Natural Language Processing (NLP). It was quickly adapted by SketchRNN [8] in order to extend [7] to free-hand sketches.

Inverse Graphics "Inverse Graphics" is line of work that aims to estimate 3D scene parameters from raster images without supervision. Instead it predicts the input parameters of a computer graphics pipeline that can reconstruct the image. Several attempts were made [26, 14] to estimate explicit model parameters of 3D objects from raw images. A specialized case of the generic Inverse Graphics idea is to estimate parameters of 2D objects such as curves. As a recent example, an RNN based agent named SPIRAL [5] learned to draw characters in terms of

pen an brush curves. SPIRAL, however, is extremely costly due to its reliance on Policy Gradient [32] reinforcement learning training and black-box renderer.

Learning for Curves Few works have studied learning for curve generation. The recent SVG Font Generator [18] trains an excellent font embedding with a recurrent vector font image generator. However it is trained with supervision rather than inverse graphics, and limited to the more structured domain of font images. Other attempts [16] also use supervised learning on synthetic data, rather than unsupervised learning on real human sketches as we consider here.

3 Methodology

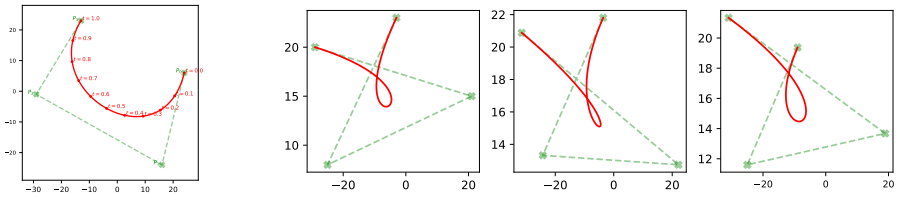
Background: Conventional Sketch representation and Generation A common format [8] for a digitally acquired sketch $\mathcal{S}$ is as a sequence of 2-tuples, each containing a $2D$ coordinate on the canvas sampled from a continuous drawing flow and a pen-state bit denoting whether the pen touches the canvas or not.

$$\mathcal{S} = \big[(\mathbf{X}_i, q_i)\big]_{i=1}^{L} \tag{1}$$

where $\mathbf{X}_i \triangleq \big[x \;\; y\big]_i^T \in \mathbb{R}^2$, $q_i \in \{\textsc{PenUp}, \textsc{PenDown}\}$ and L is the cardinality of $\mathcal{S}$ representing the length of the sketch. The state-of-the-art sketch generator SketchRNN [8] learns a parametric Recurrent Neural Network (RNN) to model the joint distribution of coordinates and pen state as a product of conditionals, i.e. $p_{sketchrnn}(\mathcal{S};\theta) = \prod_{i=1}^{L} p\big(\mathbf{X}_i, q_i \,|\mathbf{X}_{<i}, q_{<i}; \theta\big)$, where θ is the set of parameters of the model and $\mathbf{X}_{<i}$ and $q_{<i}$ denote the list of locations and pen-state bits respectively before $\mathbf{X}_i$ and q_i.

Towards a Stroke-Level Representation We are interested in moving from such a segment-level representation toward stroke-level. To this end we modify the structure of our input data to $\bar{\mathcal{S}} \triangleq \big[\mathbf{T}_j\big]_{j=1}^{N}$, with $\mathbf{T}_j \triangleq \big[\mathbf{X}_i^{(j)}\big]_{i=1}^{N_j}$ where $\mathbf{T}_j$ is the j^{th} stroke of length $N_j \triangleq |\mathbf{T}_j|$ segregated from the sketch by following the pen-state bit, and consequently $\sum_{j=1}^{N} N_j = L$.

Towards a Stroke-Level Generative Model Existing generative sketch models [8, 30] generate a segment at each iteration. Given a stroke-segmented training set $\bar{\mathcal{S}}$, we would like to train a generative model analogous to SketchRNN. That is, to model the distribution over possible sketches with a parametric model $p_{model}(\bar{\mathcal{S}};\theta)$ and that approximates the original data distribution $p_{data}(\bar{\mathcal{S}})$. Different sketches having different lengths N makes this problem suitable for Recurrent Neural Networks (RNN). One could model the probability of a sketch as a product of the probabilities of individual strokes $\mathbf{T}_j$ conditioned on all its previously seen strokes $\mathbf{T}_{<j}$ and parameterized by set of parameters θ as $p_{model}(\bar{\mathcal{S}};\theta) = \prod_j p(\mathbf{T}_j|\mathbf{T}_{<j};\theta)$. However, a problem with such an approach is that the individual strokes $\mathbf{T}_j$ are of varying length which would require a hierarchical model where $p(\mathbf{T}_j|\cdot)$ is again modeled as a sequence. So we instead propose to learn fixed length embedding $\mathbf{e}_j \triangleq \mathbf{e}(\mathbf{T}_j) \in \mathbb{R}^d$ for any stroke $\mathbf{T}_j$



(a) (b)

Fig. 2: (a) An example of Bézier curve of degree $n = 3$ with $n+1$ control points. (b) Bézier curves with Gaussian noise ($\mu = \mathbf{0}, \Sigma = 5I_2$) added to control points produce similar curves in image space.

and corresponding non-parametric decoder $\mathbf{d}(\cdot)$ such that $\mathbf{T}_j \approx \mathbf{d}(\mathbf{e}_j)$. We then model the encoded sketch $\mathbf{e}(\bar{\mathcal{S}}) \triangleq \{\mathbf{e}_j\}_{j=1}^{N}$ as

$$p_{model}(\mathbf{e}(\bar{\mathcal{S}}); \theta) = \prod_{j=1}^{N} p(\mathbf{e}_j | \mathbf{e}_{<j}; \theta) \tag{2}$$

where the individual conditionals are typically one or more mixtures of Gaussians (GMMs) and where the raw sketch can be rendered at any point by the decoder. In order to sample a new sketch from the model, we sample each j^{th} stroke from $p(\mathbf{e}_j | \mathbf{e}_{<j}; \theta)$ and render it as $\mathbf{d}(\mathbf{e}_j)$.

A natural choice for the embedding $\mathbf{e}(\cdot)$ could be an encoder RNN trained as part of a Sequence-to-Sequence autoencoder [31]. However, We take a different approach and propose a novel inverse-graphics based encoder-decoder framework $\mathbf{T} \approx \mathbf{d}(\mathbf{e}(\mathbf{T}))$ where our neural encoder $\mathbf{e}(\cdot)$ produces an *interpretable* representation because it must decode through a white-box Bézier renderer $\mathbf{d}(\cdot)$.

3.1 Stroke embedding: BézierEncoder

To train our parametric stroke embedding with an inverse graphics strategy, we must first define a differentiable 'graphics decoder' which will be later used to train our neural encoder to map human strokes to Bézier curves.

Inverse Graphics Decoder Bézier curves, used heavily in computer graphics, are smooth curves representable in a closed functional form parameterized by a sequence of $n+1$ anchor coordinates $\mathbf{P} \triangleq \begin{bmatrix} P_x & P_y \end{bmatrix}^T \in \mathbb{R}^2$ termed *control points*. A degree n Bézier curve with control points $\begin{bmatrix} \mathbf{P}_0, \mathbf{P}_1, \cdots \mathbf{P}_n \end{bmatrix}$ is represented as

$$\mathbf{C}(t; \{\mathbf{P}_i\}) = \sum_{i=0}^{n} \mathcal{B}_{i,n}(t) \cdot \mathbf{P}_i \tag{3}$$

where $t \in [0, 1]$ is the *parameter* of the curve, $\mathcal{B}_{i,n}(t) \triangleq \binom{n}{i} t^i (1-t)^{n-i}$ is the Bernstein Basis Polynomial in t and $\mathbf{C}(t) \triangleq \begin{bmatrix} \mathrm{C}_x(t) & \mathrm{C}_y(t) \end{bmatrix}^T \in \mathbb{R}^2$ denotes a point

on the curve at $t = t$. As t assumes values $0 \rightarrow 1$, the curve starts from $\mathbf{P}_0$ and ends at $\mathbf{P}_n$ and the control points $\left[\mathbf{P}_1, \cdots, \mathbf{P}_{n-1}\right]$ control the trajectory of the curve, as illustrated in Fig. 2(a). We further use $\mathcal{P}^n \triangleq \left[P_{x0}, P_{y_0}, \cdots, P_{xn}, P_{y_n}\right] \in \mathbb{R}^{2(n+1)}$ to denote elements (curves) in the continuous space of $n+1$ control points. The decoder function $\mathbf{d} : \mathcal{P} \rightarrow \mathbf{T}$ can be trivially realized by Eq. 3 with the set of t-values chosen as per resolution requirement.

We now denote $(\mathbf{T}, \mathcal{P})$ as an arbitrary stroke and its Bézier representation, where we have dropped the subscript j and superscript n for notational brevity. Using $\mathcal{P}$ as an embedding space for $\mathbf{T}$ leads to an extremely useful and key property: Given a choice of n, two similar points in $\mathcal{P}$ space correspond to similar strokes in $\mathbf{T}$ space. As a consequence, we can sample from the conditionals in Eq. 2 to generate variations of a stroke.

Property 1. Given a $(\mathbf{T}, \mathcal{P})$ pair where $\mathbf{T} = \mathbf{d}(\mathcal{P})$ and sample $\widehat{\mathcal{P}} \sim \mathcal{N}(\mathcal{P}, \sigma)$, then the decoded $\widehat{\mathbf{T}} = \mathbf{d}(\widehat{\mathcal{P}})$ is distributed as $\mathcal{N}(\mathbf{T}, \sigma')$.

Proof. Refer to Appendix A in the supplementary document for the proof. Illustrative examples are given in Fig. 2(b).

A stroke to Bézier encoder We wish to learn an embedding function $\mathbf{e}(\cdot)$ that will map a given stroke $\mathbf{T}$ to its best fit Bézier representation $\mathcal{P}$. Due to the variable length of strokes $\mathbf{T}$, we model BézierEncoder with a bi-directional RNN, with forward and backward states $\overrightarrow{\mathbf{s}_i}, \overleftarrow{\mathbf{s}_i} \in \mathbb{R}^h$ at time-step i as

$$\left[\overrightarrow{\mathbf{s}_i}, \overleftarrow{\mathbf{s}_i}\right] = \text{BiRNN}(\mathbf{X}_{i-1}, \mathbf{s}_{i-1}; \theta) \tag{4}$$

However, unlike regular encoder RNNs, we further transform the last hidden state to get a Bézier curve representation

$$\mathcal{P} = \mathbf{W}_{\mathcal{P}} \left[\overrightarrow{\mathbf{s}}_{end}; \overleftarrow{\mathbf{s}}_{end}\right] \tag{5}$$

where the '*end*' subscript denotes the state of the RNN at last time-step, [;] denotes the concatenation operator and $\mathbf{W}_{\mathcal{P}} \in \mathbb{R}^{2(n+1)\times 2h}$.

The formulation so far enables extracting a curve $\mathcal{P}$ from data $\mathbf{T}$. However, while $\mathcal{P}$ is now a sufficient representation to decode the Bézier by means of Eq. 3, we do not have sufficient information to compute a reconstruction loss like $\|\mathbf{T} - \mathbf{d}(\mathbf{e}(\mathbf{T}))\|$ because we lack the association between input coordinates $\mathbf{X}_i$ and interpolation parameters t_i. This is where many classic Bézier fitting techniques [17, 35] resort to slow alternating optimization techniques.

We take a different approach and ask our encoder to also predict the corresponding interpolation parameter t_i for each input point $\mathbf{X}_i$. In order to make valid predictions for t we note the properties it requires due to its role in Bézier curves generation: **1.** $0 \leqslant \widehat{t}_i \leqslant 1$ (by definition of Bézier curve). **2.** $\widehat{t}_i \leqslant \widehat{t}_{i+1}$ (due to sequential nature of $\mathbf{X}_i$). Apart from these, we impose another property without any lose of generality: **3.** $t_1 = 0$ and $t_{end} = 1$ (this will make $\mathbf{X_1}$ and $\mathbf{X}_{end}$ coincide with $\mathbf{P}_0$ and $\mathbf{P}_n$ respectively). Please refer to the experiment section for an implementation trick to do so.

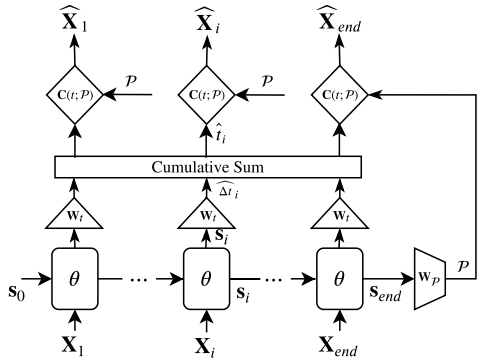


Fig. 3: Inverse graphics training of our BézierEncoder architecture for model-based single-pass stroke $[\mathbf{X}_i]$ to Bézier $\mathcal{P}$ mapping.

To enable our encoder to meet these requirements above, we do not compute t_is directly, but instead compute increments $\Delta t_i \triangleq t_i - t_{i-1}$ (with $t_0 \triangleq 0$) from $\left[\overrightarrow{\mathbf{s}_i}; \overleftarrow{\mathbf{s}_i}\right]$ at every step i. The t_i-values can then be easily computed as a cumulative sum of all Δt_i up to i. Thus, the second path of our encoder predicts

$$\widehat{t_i} = \sum_{i'=1}^{i} \widehat{\Delta t}_{i'}, \text{ with } \widehat{\Delta t}_i = \textsc{Softmax}_i(\mathbf{W}_t \cdot \left[\overrightarrow{\mathbf{s}_i}; \overleftarrow{\mathbf{s}_i}\right]). \tag{6}$$

The usage of SOFTMAX() enforces all three requirements stated above.

To summarize: Our full architecture, as shown in Figure 3 thus has two pathways: A Bézier embedding pathway that predicts the curve $\mathcal{P}$ for the entire stroke input $\mathbf{T}$ and an interpolation parameter pathway that further predicts the estimated curve parameter $\widehat{t_i}$ for each input point $\mathbf{X}_i$ in $\mathbf{T}$. Given the $(\mathbf{X}_i, \widehat{t_i})$ pairs and $\mathcal{P}$ predicted by our encoder, we can now train our model with the following reconstruction loss:

$$\mathcal{L}(\theta, \mathbf{W}_\mathcal{P}, \mathbf{W}_t) \triangleq \sum_i \left\| \mathcal{C}(\widehat{t_i}, \mathcal{P}) - \mathbf{X}_i \right\|^2 \tag{7}$$

which is optimized w.r.t. encoder parameters $\{\theta, \mathbf{W}_\mathcal{P}, \mathbf{W}_t\}$ by SGD. Once trained, we can compute the best-fit Bézier for any stroke using Eq. 5, which provides a feed-forward single pass solution to a typically alternating optimization.

A Multi-Degree Representation Extension To add more flexibility, we can extend this basic building block to learn a multi-degree representation of a given stroke $\mathbf{T}$. In order to do so, we encode the stroke using the the same RNN in Eq. 4 parameterized by θ but use a set of different $\mathbf{W}_{\mathcal{P}^n}$ and $\mathbf{W}_t^n$ for a predefined range of degree $n \in [n_{min}, \cdots, n_{max}]$ to predict Bézier representations of different degrees along with their corresponding t_i^n-values.

$$\widehat{t}_i^n = \sum_{i'=1}^{i} \widehat{\Delta t}_{i'}^n, \text{ with } \widehat{\Delta t}_i^n = \textsc{Softmax}_i(\mathbf{W}_t^n \cdot [\overrightarrow{\mathbf{s}_i}; \overleftarrow{\mathbf{s}_i}]) \text{ and } \mathcal{P}^n = \mathbf{W}_{\mathcal{P}^n} [\overrightarrow{\mathbf{s}}_{end}; \overleftarrow{\mathbf{s}}_{end}] \tag{8}$$

The total loss is now the sum of losses at every order n:

$$\mathcal{L}_{total} \triangleq \sum_{n=n_{min}}^{n_{max}} \mathcal{L}_n, \text{ with } \mathcal{L}_n(\theta, \mathbf{W}_{\mathcal{P}^n}, \mathbf{W}_t^n) \triangleq \sum_i \left\| \mathcal{C}(\widehat{t}_i^n, \mathcal{P}^n) - \mathbf{X}_i \right\|^2 \tag{9}$$

Inference in this model can now predict a *set* of Bézier representations for different degrees, where higher order curves fit the data better at the cost of more control points. The preferred order can then be chosen manually according to user requirement, or automatically by heuristic. An effective heuristics is to evaluate the loss $\mathcal{L}_n$ for all n and choose the smallest n for which $\mathcal{L}^n \leq L_{tolerance}$.

Smoothness Regularizer Our training objectives Eq. 7 or Eq. 9 may lead to overfitting in the domain of Bézier curves during encoder learning. To avoid this we add a smoothness regularizer (with regularization strength β) that prefers sequential control points to be nearby. Specifically, we add $\beta \cdot \mathcal{R}_n$ with $\mathcal{L}_n$ for each n, where $\mathcal{R}_n(\mathcal{P}^n) \triangleq \sum_{i=1}^{n} \|\mathbf{P}_{i+1} - \mathbf{P}_i\|_2^2$.

3.2 Sketch generation: BézierSketch

We next leverage our choice of Bézier representation space, and encoding model $\mathcal{P} = \mathbf{e}(\cdot)$ to define two alternative vector graphic generative models for sketches.

Control Point mode Given a sketch as a sequence of stroke embeddings $\{\mathcal{P}_j\}_{i=1}^N$ obtained from the raw input strokes as $\mathcal{P} = \mathbf{e}(\mathbf{T})$, we can modify the original data structure in Eq. 1 and substitute the set of absolute co-ordinates of every stroke by the set of control points of its Bézier representation. The modified sketch $\mathcal{S}_{cp}$ would be

$$\mathcal{S}_{cp} = \left[\left(\mathbf{P}_0^{(j)}, q_0^{(j)} \right), \cdots, \left(\mathbf{P}_i^{(j)}, q_i^{(j)} \right), \cdots, \left(\mathbf{P}_{n_j}^{(j)}, q_{n_j}^{(j)} \right) \right]_{j=1}^{N} \tag{10}$$

When encoded this way by our Bézier encoder, each sketch is represented by a relatively shorter (mostly) list of parametric control points rather than the original long list of coordinates. In this format, different strokes can have different degrees, as indicated by the use of n_j above.

Given this sequential representation of a sketch dataset, we can now train a generative sketch model. Since $\mathcal{S}_{cp}$ is structurally same as original $\mathcal{S}$ apart from its length and the interpretation of its co-ordinates, we can re-use exactly the same architecture and training procedure as SketchRNN [8]. We use a variational

sequence-to-sequence autoencoder [31] with a latent vector encoding the whole sketch. Thus one sketch is encoded first to a list of Bézier curves, and then to a latent vector in SketchRNN architecture; and decoded first to a list of curve parameters, and then rendered by the Bézier renderer. Please refer to Appendix B for a brief review of the SketchRNN architecture in the context of our problem.

Stroke mode Given a sketch $\mathcal{S}$ as set of strokes $\{\mathbf{T}_j\}_{j=1}^{N}$, we transform it as $\mathcal{S}_{st} = \{\mathcal{P}_j\}_{j=1}^{N}$ where $\mathcal{P}_j = \mathbf{e}(\mathbf{T}_j)$. We model the whole sketch using a sequence-to-sequence autoencoder, where each time-step processes one stroke represented as a fixed order Bézier curve. We use a bi-directional RNN to encode the whole sketch stroke-by-stroke. The hidden states (forward and backward) of the encoder $\overrightarrow{\mathbf{h}}_j$, $\overleftarrow{\mathbf{h}}_j$ at time-step j is given as

$$\left[\overrightarrow{\mathbf{h}}_j, \overleftarrow{\mathbf{h}}_j\right] = \text{BiRNN}(\mathcal{P}_{j-1}, \mathbf{h}_{i-1}; \Theta)$$

A latent vector $\mathbf{z} \in \mathbb{R}^{N_z}$ encoding the whole sketch is sampled using the parameters of a Gaussian distribution computed from the last hidden states

$$\mathbf{z} \sim \mathcal{N}(\mu_{\mathbf{z}}, \text{diag}(\sigma_{\mathbf{z}})), \text{ with } [\mu_{\mathbf{z}}, \sigma_{\mathbf{z}}] = f\left(\left[\overrightarrow{\mathbf{h}}_N; \overleftarrow{\mathbf{h}}_N\right]; \Theta\right)$$

An unidirectional decoder RNN is initialized using $\mathbf{z}$ and models the probability of j^{th} stroke embedding conditioned on the hidden state $\mathbf{g}_j \in \mathbb{R}^{H^d}$

$$\begin{aligned} p(\mathcal{P}_j|\mathbf{g}_j; \Theta) &= \text{GMM}\left(\mathcal{P}_j; \left\{\mu_j^m(\mathbf{g}_j), \Sigma_j^m(\mathbf{g}_j), \pi_j^m(\mathbf{g}_j)\right\}_{m=1}^{M}\right) \\ \mathbf{g}_j &= \text{DecoderRNN}([\mathcal{P}_{j-1}; \mathbf{z}], \mathbf{g}_{j-1}; \Theta) \end{aligned} \tag{11}$$

where $\left\{\mu_j^m, \Sigma_j^m, \pi_j^m\right\}$ are the parameters of the M-component GMM for the j^{th} stroke. For computational efficiency, we consider diagonal Σ_j^m and by definition $\sum_m \pi_j^m = 1$. Given a trained model, we can sample from this distribution to generate similar $\mathcal{P}_j$ which will resemble its original domain data $\mathbf{T}_j$ as guaranteed by property 1. Along with $\mathcal{P}_j$ at every step j, we also predict a stop bit $\widehat{b}_j \in [0, 1]$ denoting end of sketch which is compared against the ground-truth stop bit $b_j \triangleq \mathbb{1}_{j=N}$. The sketch generator is trained with the following objective function

$$\begin{aligned} \mathcal{L}(\{\mathcal{P}_j\}_{i=1}^{N}; \Theta) = \Bigg[&-\frac{1}{N_{max}} \sum_{j=1}^{N} \log \text{GMM}\left(\mathcal{P}_j|\left\{\mu_j^m, \Sigma_j^m, \pi_j^m\right\}_{m=1}^{M}; \Theta\right) \\ &-\frac{1}{N_{max}} \sum_{j=1}^{N} b_j \log \widehat{b}_j \Bigg] - \frac{1}{2N_z} \sum_{i=1}^{N_z} \left(1 + \sigma_{\mathbf{z}}^i - \mu_{\mathbf{z}}^i - exp(\sigma_{\mathbf{z}}^i)\right) \end{aligned} \tag{12}$$

The first two terms of $\mathcal{L}$ are the log-likelihood of a sequence $\{\mathcal{P}_j\}_{i=1}^{N}$ under the model and the loss due to the stop bit respectively. The third term denotes the KL-divergence loss for imposing a Gaussian prior on the latent code $\mathbf{z}$. The diagonal entries of Σ_j^m have been raised by $exp(\cdot)$ to make them non-negative and $\textsc{Softmax}(\cdot)$ has been used to ensure $\sum_m \pi_j^m = 1$.

4 Experiments & Results

Dataset *Quick, Draw!* is a large sketch dataset [8] collected as a part of an online game to draw a given category within a time-limit, in which thousands of people around the world participated. Due to the problem definition and structure of data used by our framework (see Eq.1), *Quick, Draw!* is the most suitable dataset to validate it. Different versions of the dataset use different sampling rates at which the sketches are stored as point sequences. SketchRNN is known to work well only on data with lower sampling rate (i.e., $\mathbb{E}_{\mathbf{T}}\big[|\mathbf{T}|\big]$ is lower) than the raw data ($\mathbb{E}_{\mathbf{T}}\big[|\mathbf{T}|\big]$ is higher) recorded. Due to fixed length of Bézier representations, our framework can adapt to data with both high and low sampling rates without any modification. Although our method is generalizable across all categories, we experimented with few categories to validate our claims.

Our framework has two main components: **1.** Embedding each stroke into its Bézier representation. **2.** Training a generative model with the encoded sketches either in *control point mode* or *stroke mode*. As our BézierEncoder is a key contribution, we validate this in isolation, before comparing our whole BézierSketch framework to SketchRNN [8].

4.1 Stroke Embedding Experiments

Implementation Details We created a dataset of all strokes from all sketches in a category of *Quick, Draw!* in order to train the stroke embedding model described in Section 3.1. We adopted some tricks that made the training and representation more efficient in practice. We normalized all strokes to start from the origin (i.e., $\mathbf{X}_1 = [0, 0]^T$). Furthermore, we assumed that the first control point $\mathbf{P}_0$ of a Bézier representation is always aligned to the first absolute coordinate of the stroke (i.e., $\mathbf{X}_1 = \mathbf{P}_0$). Given these design choices, we can ignore the first control point (fixing it to origin) and only predict successive differences of control points (i.e., $\Delta\mathbf{P}_1 \triangleq \mathbf{P}_1 - \mathbf{P}_0$, $\Delta\mathbf{P}_2 \triangleq \mathbf{P}_2 - \mathbf{P}_1$ and so on) and then decode $\mathbf{P}_i$ as $\mathbf{P}_i = \sum_{i'=1}^{i} \Delta\mathbf{P}_{i'}$ while evaluating the loss in Eq. 7. We chose the hidden state dimension to be $h = 256$ and $n_{min} = 3, n_{max} = 9$ for learning multi-degree Bézier representation. To exclude over complicated strokes, we apply some heuristics to split a stroke into two or more. Specifically, we split a stroke into multiple parts based on two criteria: 1. Every part is within a maximum length and 2. Every part has only one sharp bend (determined by computing its curvature at a given point). We set the regularizer weight $\beta = 10^{-3}$.

Results We first qualitatively demonstrate the results of inferring Bézier representations of input strokes. Fig. 4(top left) shows fitting results for various curve orders (columns) – showing variable amounts of detail being captured at different orders. It also shows fitting examples at both low (above) and high (below) sampling rates – confirming that our encoder can adapt to both.

We next qualitatively illustrate the training dynamics of our model via the fit estimated as training progresses. The results in Fig. 4(middle) show the estimated fit during training in terms of Bézier curve (red) and control points

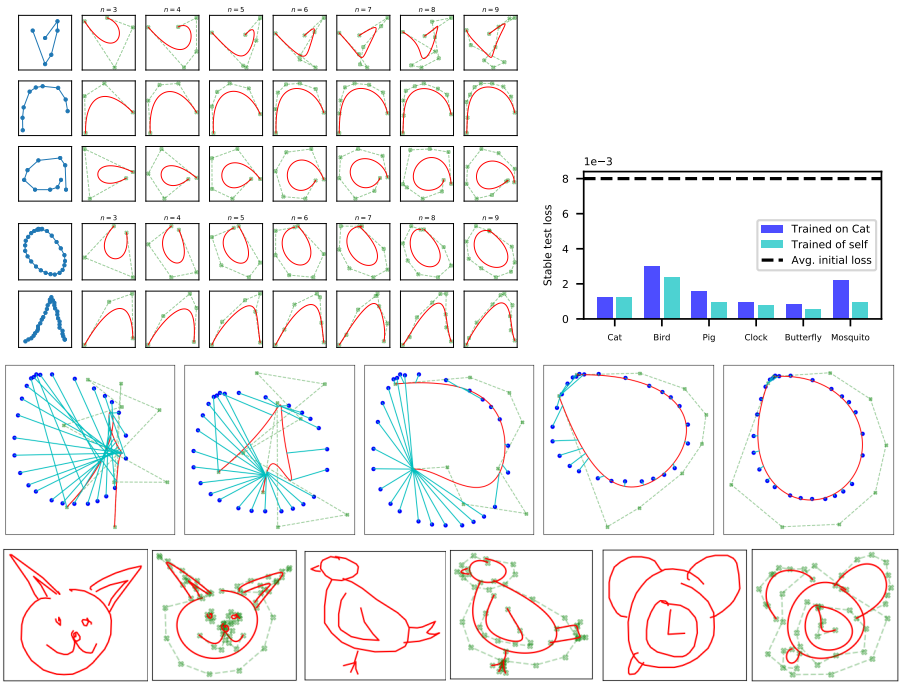


Fig. 4: Evaluating our BézierEncoder. (Top left) Learned representations of multi-degree Bézier stroke embedding. Top and bottom rows contain moderate and high-sampling rate respectively. (Top right) Test loss for various categories when trained on same category vs "Cat", demonstrating transferability of the encoder. (Middle) Visualising training dynamics. Blue: Stroke to fit. Red and Green: Bézier curve and control points. Cyan: Estimated point correspondence. (Bottom) Examples of full sketches and their learned Bézier representation.

(green) for a stroke defined by (blue) points. Recall that our encoder also predicts the interpolation parameters t that match each input point to a location on the curve. These correspondences are indicated in (cyan). Clearly both the fit and the estimated correspondences improve with training iterations. Refer to Appendix C in the supplementary document for similar visualization of more samples.

Given that our training data is grouped into categories, we next verify that our encoder indeed learns a generic Bèzier embedding, and is not overfitted to a specific category. Specifically, we compare the test loss for reconstructing data of each category when the encoder is trained on the same category as testing vs trained vs a disjoint category to testing. The results in Fig. 4(top right) shows that the embedding generalizes quite well to categories it is not trained on.

Finally, Fig. 4(bottom) shows examples of full sketches encoded by our encoder, and then decoded as Béziers. We can see that the encoded sketches reflect the input, but are smoother and cleaner.

4.2 Sketch generation Experiments

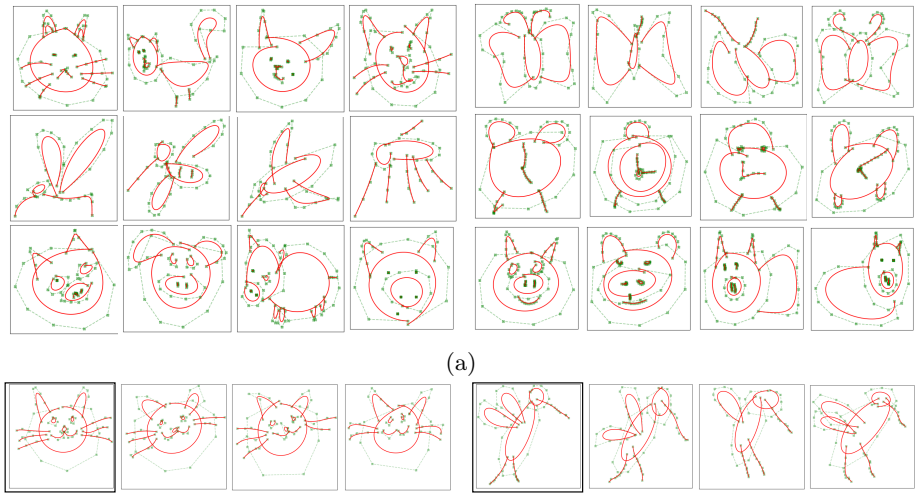
Setup In control point mode, a fully trained multi-degree embedding model is used to restructure all sketches in our dataset as $\mathcal{S}_{cp}$. We set $\mathcal{L}_{tolerance} = 10^{-3}$ to select the best n. We then train a SketchRNN-like model [8] using the restructured data. As data augmentation, we added 2D standard normal noise at all control points. Sampling from the latent space and decoding it by the decoder will generate sequence of control points and stroke/sketch ending bits. Treating one entire stroke as a set of control points, we can then draw it on a canvas using Eq. 3 with any required level of granularity.

In stroke mode, we encode each stroke with a fixed degree of $n = 9$. Very similar to *control point mode*, we use a Bi-LSTM to encode the whole sketch stroke-by-stroke and extract N_z dimensional latent vector. By conditioning on the latent vector, the decoder produces Bézier representation $\mathcal{P}$ of one stroke at each time-step. Thus, the length of a sketch coincides with the number of strokes present in the sketch. At each step of the decoder, we sample one stroke from $p(\mathcal{P}_j|\mathbf{g}_j, \Theta)$ which is modeled as a GMM with $M = 10$ mixture components. However, unlike the control point mode and its corresponding SketchRNN-like architecture, we do not use correlation parameter in the constituent Gaussians. This design choice makes the individual dimensions of the Gaussians independent, sampling from which is justified given property. 1. Apart from $\mathcal{P}_j$, we predict one more quantity in practice: the start location $\mathbf{v}_j \triangleq (v_x, v_y)_j^T$ of the stroke w.r.t the whole sketch. The need for $\mathbf{v}_j$ arises due to the practical consideration of relocating the start of each individual stroke at the origin while encoding them.

Results Qualitative results of generated unconditional sketch samples from both our model variants are shown in Fig. 5(a). We can see that, similarly to SketchRNN, BézierSketch generates diverse and plausible samples. However, uniquely our samples are high-resolution vector graphic sketches. Fig. 5(b) also shows examples of conditional samples where the right group of three images are samples conditioned on the left sketch encoding.

The use of Bézier curves as stroke representation reduces the average length of a given stroke's representation significantly and as a direct consequence, the description length for whole sketches as well. In Fig. 6, we compare the length histograms of original data and its Bézier representation both on stroke and sketch level, confirming that Béziers are systematically shorter (left). This is the same for strokes and sketches sampled by vanilla and SketchRNN and BézierSketch respectively (right).

This property of shorter representations for any given sketch means that our generator should have an advantage modeling longer sketches compared to vanilla SketchRNN since it only needs to model shorter sequences. To evaluate this, we use a modified Fréchet Inception Distance (FID) [9] score to compare the generated samples from both models. We first trained both our generator model and SketchRNN on the entire dataset (of each category). We then create a subset of sketches whose original length is $l \pm 20$ and use them to generate samples.



(a)

(b)

Fig. 5: Qualitatively evaluating BézierSketch. (a) Samples drawn unconditionally in control point mode (left half) and stroke mode (right half). (b) Sketch samples generated by conditioning on the first sketch (double bordered) in each set.

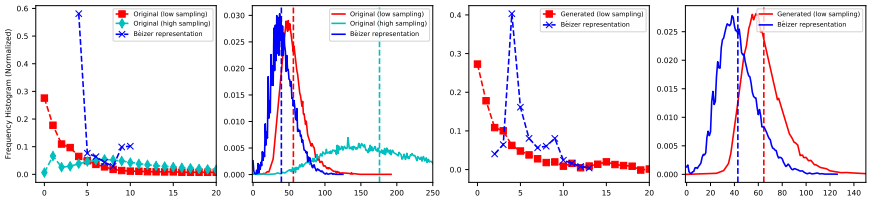


Fig. 6: Stroke/Sketch Length histogram for original data (left) and generated samples (right). Bézier encodings are shorter sequences than the raw data.

All original and generated samples are rendered on a canvas and projected down to a concise feature vector using pre-trained Sketch-a-Net 2.0 [33] classifier. We compute the empirical mean and covariance of both real samples and generated samples as (μ_r, Σ_r) and (μ_g, Σ_g) and then estimate modified FID as:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$$

The results in Fig. 7 plots the modified FID score with increasing length value l for both SketchRNN and our model on each category of sketches. We can see that our model leads to improved (lower) FID score, especially for longer sketches. This is illustrated qualitatively in Fig. 7, where we can see that for longer sketches, our framework produces much more reliable reconstruction than QuickDraw, which fails to make reasonable reconstruction in these cases.

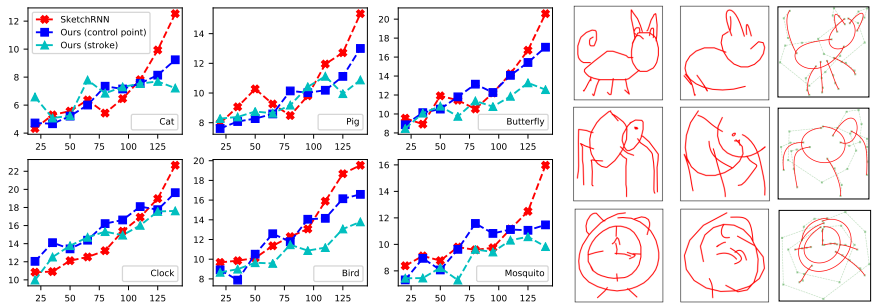


Fig. 7: Left: FID score (↓) vs length of sketch shows the effectiveness of our generative model on longer sketches. Right: Qualitative samples of long sketches. Three columns denote the original sketch, SketchRNN and our BézierSketch.

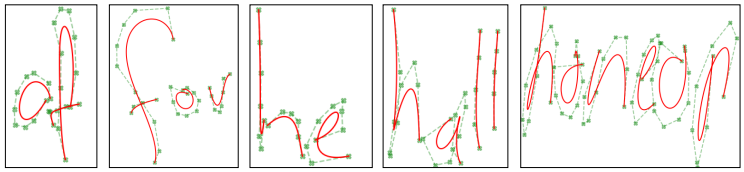


Fig. 8: Unconditionally generating handwritten words from the IAM database.

Other applications Although crafted with sketches in mind, our framework can be adapted to other applications like handwriting generation (in line with the work of [7]) with little to no modification. In fact, any 2D sequence data with two-level hierarchical representation (e.g., stroke and sketch) can be modeled using the same framework. Online handwritten characters are composed of relatively short strokes which we model with Bézier curves. We use the online handwritten sentences from the IAM handwriting database [19], embed the constituent strokes with our Bézier representation and train our generative model for words. Fig. 8 shows qualitative samples from our resulting word generator.

5 Conclusions

In this paper we presented an inverse graphics approach to training an efficient model-based single-pass stroke-to-Bézier encoder via reconstruction through a Bézier decoder. Such approach surpasses the conventional fitting-based methods in terms of quality and efficiency. Furthermore, this enabled us to advance generative sketch models by generating sketches as sequences of parameterized curves rather than pixels, leading to arbitrary-resolution scalable vector graphic samples. This new representation also enables better generation of longer sketches compared to existing state of the art. In future work we will investigate extending to more complex parameterized curves such as B-splines, and developing an encoder to predict curves from rasterized images directly.

References

1. Bishop, C.M.: Mixture density networks. Tech. rep., Aston University (1994)
2. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. In: CoNLL (2016)
3. De Boor, C., De Boor, C., Mathématicien, E.U., De Boor, C., De Boor, C.: A practical guide to splines, vol. 27. Springer-Verlag New York (1978)
4. Dey, S., Riba, P., Dutta, A., Llados, J., Song, Y.Z.: Doodle to search: Practical zero-shot sketch-based image retrieval. In: CVPR (2019)
5. Ganin, Y., Kulkarni, T., Babuschkin, I., Eslami, S.M.A., Vinyals, O.: Synthesizing programs for images using reinforced adversarial learning. In: ICML (2018)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
7. Graves, A.: Generating sequences with recurrent neural networks. CoRR **abs/1308.0850** (2013)
8. Ha, D., Eck, D.: A neural representation of sketch drawings. In: ICLR (2018)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS (2017)
10. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. ICLR (2014)
13. Klare, B., Li, Z., Jain, A.: Matching forensic sketches to mug shot photos. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(3), 639 –646 (march 2011)
14. Kulkarni, T.D., Whitney, W., Kohli, P., Tenenbaum, J.B.: Deep convolutional inverse graphics network. In: NIPS (2015)
15. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
16. Laube, P., Franz, M.O., Umlauf, G.: Deep learning parametrization for b-spline curve approximation. In: 2018 International Conference on 3D Vision (3DV) (2018)
17. Liu, Y., Wang, W.: A revisit to least squares orthogonal distance fitting of parametric curves and surfaces. In: GMP (2008)
18. Lopes, R.G., Ha, D., Eck, D., Shlens, J.: A learned representation for scalable vector graphics. In: ICCV (2019)
19. Marti, U.V., Bunke, H.: A full english sentence database for off-line handwriting recognition. In: ICDAR (1999)
20. Masood, A., Ejaz, S.: An efficient algorithm for robust curve fitting using cubic bezier curves. In: ICIC (2010)
21. Pang, K., Li, K., Yang, Y., Zhang, H., Hospedales, T.M., Xiang, T., Song, Y.Z.: Generalising fine-grained sketch-based image retrieval. In: CVPR (2019)
22. Plass, M., Stone, M.: Curve-fitting with piecewise parametric cubics. In: SIGGRAPH (1983)
23. Rabiner, L., Juang, B.: An introduction to hidden markov models. IEEE ASSP Magazine **3**(1), 4–16 (1986)
24. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)

25. Revow, M., Williams, C.K.I., Hinton, G.E.: Using generative models for handwritten digit recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **18**(6), 592–606 (1996)
26. Romaszko, L., Williams, C.K.I., Moreno, P., Kohli, P.: Vision-as-inverse-graphics: Obtaining a rich 3d explanation of a scene from a single image. In: ICCVW (2017)
27. Salomon, D.: Curves and surfaces for computer graphics. Springer Science & Business Media (2007)
28. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: Learning to retrieve badly drawn bunnies. In: SIGGRAPH (2016)
29. Shao, L., Zhou, H.: Curve fitting with bezier cubics. Graphical models and image processing **58**(3), 223–232 (1996)
30. Song, J., Pang, K., Song, Y., Xiang, T., Hospedales, T.M.: Learning to sketch with shortcut cycle consistency. In: CVPR (2018)
31. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: ICML (2015)
32. Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: NIPS (1999)
33. Yu, Q., Yang, Y., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.: Sketch-a-net: A deep neural network that beats humans. International Journal of Computer Vision **122**, 411–425 (2017)
34. Yu, Q., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.: Sketch-a-net that beats humans. In: BMVC (2015)
35. Zheng, W., Bo, P., Liu, Y., Wang, W.: Fast b-spline curve fitting by l-bfgs. Computer Aided Geometric Design **29**(7), 448–462 (2012)