

# Generating Handwriting via Decoupled Style Descriptors

Atsunobu Kotani , Stefanie Tellex , and James Tompkin 

Brown University

**Abstract.** Representing a space of handwriting stroke styles includes the challenge of representing both the style of each character and the overall style of the human writer. Existing VRNN approaches to representing handwriting often do not distinguish between these different style components, which can reduce model capability. Instead, we introduce the Decoupled Style Descriptor (DSD) model for handwriting, which factors both character- and writer-level styles and allows our model to represent an overall greater space of styles. This approach also increases flexibility: given a few examples, we can generate handwriting in new writer styles, and also now generate handwriting of new characters across writer styles. In experiments, our generated results were preferred over a state of the art baseline method 88% of the time, and in a writer identification task on 20 held-out writers, our DSDs achieved 89.38% accuracy from a single sample word. Overall, DSDs allows us to improve both the quality and flexibility over existing handwriting stroke generation approaches.

## 1 Introduction

Producing computational models of handwriting is a deeply *human* and *personal* topic—most people can write, and each writer has a unique style to their script. Capturing these styles flexibly and accurately is important as it determines the space of descriptive expression of the model; in turn, these models define the usefulness of our recognition and generation applications. For deep-learning-based models, our concern is how to architecture the neural network such that we can represent the underlying stroke characteristics of the styles of writing.

Challenges in handwriting representation include reproducing fine detail, generating unseen characters, enabling style interpolation and transfer, and using human-labeled training data efficiently. Across these, one foundational problem is how to succinctly represent both the style variation of each character and the overall style of the human writer—to capture both the variation within an ‘h’ letterform and the overall consistency with other letterform for each writer.

As handwriting strokes can be modeled as a sequence of points over time, supervised deep learning methods to handwriting representation can use recurrent neural networks (RNNs) [17,2]. This allows consistent capture of style features that are distant in time and, with the use of variational RNNs (VRNNs), allows the diverse generation of handwriting by drawing from modeled distributions. However, the approach of treating handwriting style as a ‘unified’ property of a

sequence can limit the representation of both character- and writer-level features. This includes specific character details being averaged out to maintain overall writer style, and an reduced representation space of writing styles.

Instead, we explicitly represent 1) writer-, 2) character- and 3) writer-character-level style variations within an RNN model. We introduce a method of Decoupled Style Descriptors (DSD) that models style variations such that character style can still depend on writer style. Given a database of handwriting strokes as timestamped sequences of points with character string labels [38], we learn a representation that encodes three key factors: writer-independent character representations ( $\mathbf{C}_h$  for character  $h$ ,  $\mathbf{C}_{his}$  for the word  $his$ ), writer-dependent character-string style descriptors ( $\mathbf{w}_h$  for character  $h$ ,  $\mathbf{w}_{his}$  for the word  $his$ ), and writer-dependent global style descriptors ( $\mathbf{w}$  per writer). This allows new sequence generation for existing writers (via new  $\mathbf{w}_{she}$ ), new writer generation via style transfer and interpolation (via new  $\mathbf{w}$ ), and new character generation in the style of existing writers (via new  $\mathbf{C}_2$ , from only a few samples of character 2 from *any* writer). Further, our method helps to improve generation quality as more samples are provided for projection, rather than tending towards average letterforms in existing VRNN models.

In a qualitative user study, our model’s generations were preferred 88% of the time over an existing baseline [2]. For writer classification tasks on a held-out 20-way test, our model achieves accuracy of 89.38% from a single word sample, and 99.70% from 50 word-level samples. In summary, we contribute:

- Decoupled Style Descriptors as a way to represent latent style information;
- An architecture with DSDs to model handwriting, with demonstration applications in generation, recognition, and new character adaptation; and
- A new database—BRUSH (BRown University Stylus Handwriting)—of hand-written digital strokes in the Latin alphabet, which includes 170 writers, 86 characters, 488 common words written by all writers, and 3668 rarer words written across writers.

Our dataset, code, and model will be open source at <http://dsd.cs.brown.edu>.

## 2 Related Work

Handwriting modeling methods either handle images, which capture writing appearance, or handle the underlying strokes collected via digital pens. Each may be online, where observation happens along with writing, or offline. Offline methods support historical document analysis, but cannot capture the motion of writing. We consider an online stroke-based approach, which avoids the stroke extraction problem and allows us to focus on modelling style variation. Work also exists in the separate problem of typeface generation [12,5,37,26,47].

*General style transfer methods.* Current state-of-the-art style transfer works use a part of the encoded reference sample as a style component, e.g., the output of a CNN encoder for 2D images [28,34], or the last output of an LSTM for speech audio [40]. These can be mixed to allocate parts of a conditioning style

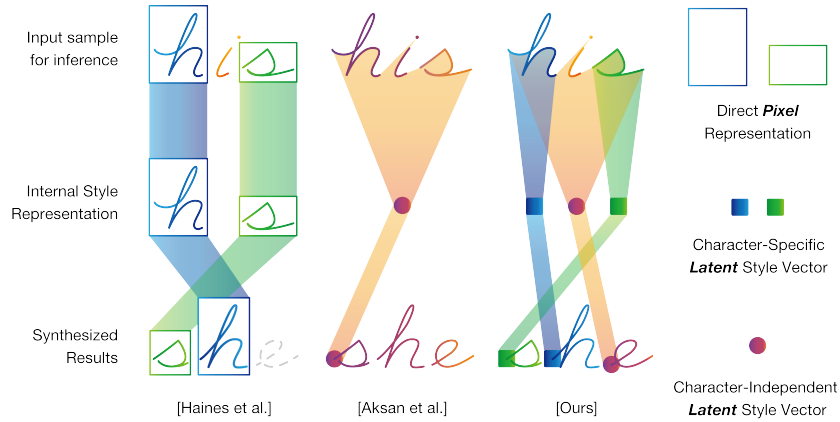


Fig. 1: Illustrating synthesis approaches. Given test sample *his* for reference, we wish to generate *she* in the same style. Left: Pixels of *h* and *s* are copied from input with a slight modification [21]; however, this fails to synthesize *e* as it is missing in the reference. Middle: A global latent writer style is inferred from *his* and used as the initial state for LSTM generation [2]. Right: Our approach infers both character and writer style vectors to improve quality and flexibility.

vector to disentangled variation [27]. Common style representations often cannot capture small details, with neural networks implicitly filtering out this content information, because the representations fail to structurally decouple style from content in the style reference source. Other approaches [30,44] tackle this problem by making neural networks predict parameters for a transformation model (an idea that originates from neuroevolution [42,20]); our **C** prediction is related.

*Recent image-based offline methods.* Haines et al. produced a system to synthesize new sentences in a specific style inferred from source images [21]. Some human intervention is needed during character segmentation, and the model can only recreate characters that were in the source images. Alonso et al. addressed the labeling issue with a GAN-based approach [16,3]; however, their model presents an image quality trade-off and struggles to generate new characters. There are also studies on typeface generation from few reference data [4,43]: Baluja generates typefaces for Latin alphabets [6], and Lian et al. for Chinese [36]. Our method does not capture writing implement appearance, but does provides underlying stroke generation and synthesizes new characters from few examples.

*Stroke-based online methods.* Deep learning methods, such as Graves’ work, train RNN models conditioned on target characters [17,13,46]. The intra-variance of a writer’s style was achieved with Mixture Density Networks (MDN) as the final synthesis layer [10]. Berio et al. use recurrent-MDN for graffiti style transfer [9]. However, these methods cannot learn to represent handwriting styles per writer, and so cannot perform writer style transfer.

Table 1: Property comparison of state-of-the-art handwriting generation models.

Method	Style transfer?	No human segmentation?	Infinite variations?	Synthesize missing samples?	Benefit from more samples?	Smooth interpolation?	Learn new characters?
Graves (2013)	No	Yes	Yes	No	No	No	No
Berio et al. (2017)	Yes	Yes	Yes	No	No	Sort of	No
Haines et al. (2017)	Yes	No	Sort of	No	Yes	No	No
Aksan et al. (2018)	Yes	Sort of	Yes	Yes	No	Sort of	No
Ours	Yes	Yes	Yes	Yes	Yes	Yes	Yes

State-of-the-art models can generate characters in an inferred style [2]. Aksan et al.’s DeepWriting model uses Variational Recurrent Neural Networks (VRNN) [15] and assumes a latent vector  $z$  that controls writer handwriting style. Across writers, this method tends to average out specific styles and so reduces detail. Further, while sample efficient, VRNN models have trouble exploiting an abundance of inference samples because the style representation is only the last hidden state of an LSTM. We avoid this limitation by extracting character-dependent style vectors from samples and querying them as needed in generation.

*Sequence methods beyond handwriting.* Learning-based architectures for sequences were popularized in machine translation [14], where the core idea is to encode sequential inputs into a fixed-length latent representation. Likewise, text-to-speech processing has been improved by sequence models [39,41], with extensions to style representation for speech-related tasks like speaker verification and voice conversion. Again, one approach is to use the (converted) last output of an LSTM network as a style representation [23].

Other approaches [24,25] models multiple stylistic latent variables in a hierarchical manner and introduces an approach to transfer styles within a standard VAE setting [33].

Broadly, variational RNN approaches [2,15,24] have the drawback that they are incapability of improving generation performance with more inference samples. While VRNNs are sample efficient when only a few samples are available for style inference, a system should also generate better results as more inference samples are provided (as in [21]). Our method attempts to be scalable and sample efficient through learning decoupled underlying generation factors.

We compare properties of four state of the art handwriting synthesis models (Tab. 1), and illustrate two of their different approaches (Fig. 1).

### 3 Method

*Input, preprocess, and output.* A stroke sequence  $\mathbf{x} = (p_1, \dots, p_N)$  has each  $p_t$  store the change in  $x$ - and  $y$ -axis from the previous timestep ( $\Delta x_t = x_t - x_{t-1}$ ,  $\Delta y_t = y_t - y_{t-1}$ ), and a binary termination flag for the ‘end of stroke’ ( $eos = \{0, 1\}$ ). This creates an  $(N, 3)$  matrix. A character sequence  $\mathbf{s} = (\mathbf{c}_1, \dots, \mathbf{c}_M)$  contains character vectors  $\mathbf{c}_t$  where each is a one-hot vector of length equal to the total number of characters considered. This similarly is an  $(M, Q)$  matrix.

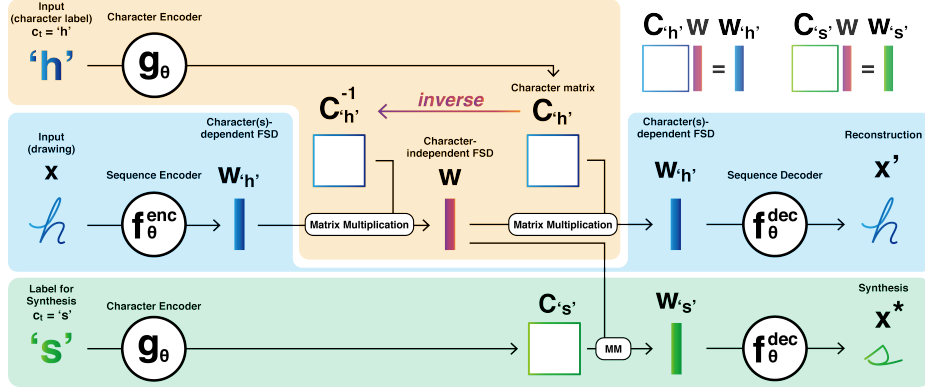


Fig. 2: High-level architecture. Circles are parametrized function approximators, and rectangles/squares are variables. *Blue region*: Encoder-decoder architecture. *Orange region*: Character-conditioned layers. *Green region*: Synthesis procedure.

The IAM dataset [38] and our stroke dataset were collected by asking participants to naturally write character sequences or words, which often produces cursive writing. As such, we must solve a segmentation problem to attribute stroke points to specific characters in  $\mathbf{s}$ . This is complex; we defer explanation to our supplemental. For now, it is sufficient to say that we use unsupervised learning to train a segmentation network  $k_\theta(\mathbf{x}, \mathbf{s})$  to map regions in  $\mathbf{x}$  to characters, and to demark ‘end of character’ labels ( $eoc = \{0, 1\}$ ) for each point.

As output, we wish to predict  $\mathbf{x}'$  comprised of  $\mathbf{p}'_t$  with: 1) coefficients for Mixture Density Networks [10] ( $\pi_t, \mu_x, \mu_y, \sigma_x, \sigma_y, \rho$ ), which provide variation in output by sampling  $\Delta x_t$  and  $\Delta y_t$  from these distributions at runtime; 2) ‘end of stroke’  $eos$  probability; and 3) ‘end of character’  $eoc$  probability. This lets us generate cursive writing when  $eos$  probability is low and  $eoc$  probability is high.

*Decoupled Style Descriptors (DSD)*. We begin with the encoder-decoder architecture proposed by Cho et al. [14] (Fig. 2, blue region). Given a supervised database  $\mathbf{x}, \mathbf{s}$  and a target string  $c_t$ , to represent handwriting style we train a parameterized encoder function  $f_\theta^{\text{enc}}$  to learn writer-dependent character-dependent latent vectors  $\mathbf{w}_{c_t}$ . Then, given  $\mathbf{w}_{c_t}$ , we simultaneously train a parameterized decoder function  $f_\theta^{\text{dec}}$  to predict the next point  $p'_t$  given all past points  $p'_{1:t-1}$ . Both encoder and decoder  $f_\theta$  are RNNs such as LSTM models:

$$p'_t = f_\theta^{\text{dec}}(p'_{1:t-1} | \mathbf{w}_{c_t}). \quad (1)$$

This method does not factor character-independent writer style; yet, we have no way of explicitly describing this property via supervision and so we must devise a construction to learn it implicitly. Thus, we add a layer of abstraction (Fig. 2, orange region) with three assumptions:

1. If two stroke sequences  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are written by the same writer, then consistency in their writing style is manifested by a character-independent writer-dependent latent vector  $\mathbf{w}$ .
2. If two character sequences  $\mathbf{s}_1$  and  $\mathbf{s}_2$  are written by different writers, then consistency in their stroke sequences is manifested by a character-dependent writer-independent latent matrix  $\mathbf{C}$ .  $\mathbf{C}$  can be estimated via a parameterized encoder function  $g_\theta$ , which is also an RNN such as an LSTM:

$$\mathbf{C}_{c_t} = g_\theta(\mathbf{s}, c_t). \quad (2)$$

3.  $\mathbf{C}_{c_t}$  instantiates a writer’s style  $\mathbf{w}$  to draw a character via  $\mathbf{w}_{c_t}$ , such that  $\mathbf{C}_{c_t}$  and  $\mathbf{w}$  are latent factors:

$$\mathbf{w}_{c_t} = \mathbf{C}_{c_t} \mathbf{w}, \quad (3)$$

$$\mathbf{w} = \mathbf{C}_{c_t}^{-1} \mathbf{w}_{c_t}. \quad (4)$$

This method assumes that  $\mathbf{C}_{c_t}$  is invertible, which we will demonstrate in Sec. 4. Intuitively, the multiplication of writer-dependent character vectors  $\mathbf{w}_{c_t}$  with the inverse of character-DSD  $\mathbf{C}_{c_t}^{-1}$  (Eq. 4) factors out character-dependent information from writer-dependent information in  $\mathbf{w}_{c_t}$  to extract a writer style representation  $\mathbf{w}$ . Likewise, Eq. 3 restores writer-dependent character  $\mathbf{w}_{c_t}$  by multiplying the writer-specific style  $\mathbf{w}$  with a relevant character-DSD  $\mathbf{C}_{c_t}$ .

We use this property in synthesis (Fig. 2, green region). Given a target character  $c_t$ , we use encoder  $g_\theta$  to generate a  $\mathbf{C}$  matrix. Then, we multiply  $\mathbf{C}_{c_t}$  by a desired writer style  $\mathbf{w}$  to generate  $\mathbf{w}_{c_t}$ . Finally, we use trained decoder  $f_\theta^{\text{dec}}$  to create a new point  $p'_t$  given previous points  $p'_{1:t-1}$ :

$$p'_t = f_\theta^{\text{dec}}(p'_{1:t-1} | \mathbf{w}_{c_t}), \text{ where } \mathbf{w}_{c_t} = \mathbf{C}_{c_t} \mathbf{w}. \quad (5)$$

*Interpreting the linear factors.* Eq. 3 states a linear relationship between  $\mathbf{C}_{c_t}$  and  $\mathbf{w}$ . This exists at the latent representation level:  $\mathbf{w}_{c_t}$  and  $\mathbf{C}_{c_t}$  are separately approximated by independent neural networks  $f_\theta^{\text{enc}}$  and  $g_\theta$ , which themselves are nonlinear function approximators [30, 44]. As  $\mathbf{C}_{c_t}$  maps a vector  $\mathbf{w}$  to another vector  $\mathbf{w}_{c_t}$ , we can consider  $\mathbf{C}_{c_t}$  to be a fully-connected neural network layer (without bias). However, unlike standard layers,  $\mathbf{C}_{c_t}$ ’s weights are not implicitly learned through backpropagation but are predicted by a neural network  $g_\theta$  in Eq. 2. A further interpretation of  $\mathbf{C}_{c_t}$  and  $\mathbf{C}_{c_t}^{-1}$  as two layers of a network is that they respectively share a set of weights and their inverse. Explicitly forming  $\mathbf{C}_{c_t}$  in this linear way makes it simple to estimate  $\mathbf{C}_{c_t}$  for *new* characters that are not in the training dataset, given few sample pairs of  $\mathbf{w}_{c_t}$  and  $\mathbf{w}$ , using standard linear least squares methods (Sec. 4).

*Mapping character and stroke sequences with  $f_\theta$  and  $g_\theta$ .* Next, we turn our attention to how we map sequences of characters and strokes within our function approximators. Consider the LSTM  $f_\theta^{\text{enc}}$ : Given a character sequence  $\mathbf{s}$  as size of  $(M, Q)$  where  $M$  is the number of characters, and a stroke sequence  $\mathbf{x}$  of size

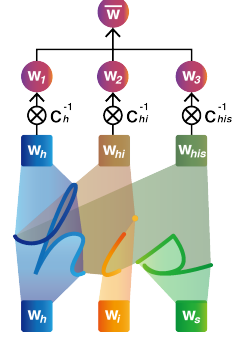
$(N, 3)$  where  $N$  is the number of points, our goal is to obtain a style vector for each character  $\mathbf{w}_{c_t}$  in that sequence. The output of our segmentation network  $k_\theta$  preprocess defines ‘end of character’ bits, and so we know at which point in  $\mathbf{x}$  that a character switch occurs, e.g., from  $h$  to  $e$  in *hello*.

First, we encode  $\mathbf{x}$  using  $f_\theta^{\text{enc}}$  to obtain a  $\mathbf{x}^*$  of size  $(N, L)$ , where  $L$  is the latent feature dimension size (we use 256). Then, from  $\mathbf{x}^*$ , we extract  $M$  vectors at these switch indices—these are our writer-dependent character-dependent DSDs  $\mathbf{w}_{c_t}$ . As  $f_\theta^{\text{enc}}$  is an LSTM, the historical sequence data up to that index is encoded within the vector at that index (Fig. 3, top). For instance, for *his*,  $\mathbf{x}^*$  at switch index 2 represents how the writer writes the first two characters *hi*, i.e.,  $\mathbf{w}_{hi}$ . We refer to these  $\mathbf{w}_{c_t}$  as ‘writer-character-DSDs’.

Likewise, LSTM  $g_\theta$  takes a character sequence  $\mathbf{s}$  of size  $(M, Q)$  and outputs an array of  $\mathbf{C}$  matrices that forms a tensor of size  $(M, L, L)$  and preserves sequential dependencies between characters: The  $i$ -th element of the tensor  $\mathbf{C}_{c_i}$  is a matrix of size  $(L, L)$ —that is, it includes information about previous characters up to and including the  $i$ -th character. Similar to  $\mathbf{x}^*$ , for *his*, the second character matrix  $\mathbf{C}_{c_2}$  contains information about the first two characters *hi*— $\mathbf{C}$  is really a character sequence matrix. Multiplying character information  $\mathbf{C}_{c_t}$  with writer style vector  $\mathbf{w}$  creates a writer-character-DSD  $\mathbf{w}_{c_t}$ .

*Estimating  $\mathbf{w}$ .* When we encode a stroke sequence  $\mathbf{x}$  that draws  $s$  characters via  $f_\theta^{\text{enc}}$ , we extract  $M$  character(s)-dependent DSDs  $\mathbf{w}_{c_t}$  (e.g.,  $\mathbf{w}_h$ ,  $\mathbf{w}_{hi}$  and  $\mathbf{w}_{his}$ , *right*). Via Eq. 4, we obtain  $M$  distinct candidates for writer-DSDs  $\mathbf{w}$ . To overcome this, for each sample, we simply take the mean to form  $\bar{\mathbf{w}}$ :

$$\bar{\mathbf{w}} = \frac{1}{M} \sum_{t=1}^M \mathbf{C}_{c_t}^{-1} \mathbf{w}_{c_t}. \quad (6)$$



*Generation approaches via  $\mathbf{w}_{c_t}$ .* Consider the synthesis task in Fig. 1: given our trained model, generate how a new writer would write *she* given a reference sample of them writing *his*. From the *his* sample, we can extract 1) segment-level writer-character-DSDs ( $\mathbf{w}_h$ ,  $\mathbf{w}_i$ ,  $\mathbf{w}_s$ ), and 2) the global  $\bar{\mathbf{w}}$ . To synthesize *she*, our model must predict three writer-character-DSDs ( $\mathbf{w}_s$ ,  $\mathbf{w}_{sh}$ ,  $\mathbf{w}_{she}$ ) as input to the decoder  $f_\theta^{\text{dec}}$ . We introduce two methods to estimate  $\mathbf{w}_{c_t}$ :

$$\text{Method } \alpha : \mathbf{w}_{c_t}^\alpha = \mathbf{C}_{c_t} \bar{\mathbf{w}} \quad (7a)$$

$$\text{Method } \beta : \mathbf{w}_{c_t}^\beta = h_\theta([\mathbf{w}_{c_1}, \dots, \mathbf{w}_{c_t}]) \quad (7b)$$

where  $h_\theta$  is an LSTM that restore dependencies between temporally-separated writer-character-DSDs as illustrated in Fig. 3, green rectangle. We train our model to reconstruct  $\mathbf{w}_{c_t}$  both ways. This allows us to use method  $\alpha$  when test reference samples do not include target characters, e.g., *his* is missing an *e* for *she*, and so we can reconstruct  $\mathbf{w}_e$  via  $\bar{\mathbf{w}}$  and  $\mathbf{C}_e$  (Fig. 3, right). It also allows us to use Method  $\beta$  when test reference samples include relevant characters that,

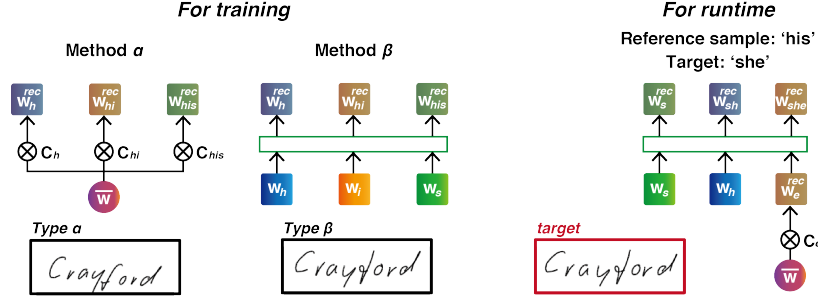


Fig. 3: Reconstruction methods to produce writer-character-DSD  $\mathbf{w}_{c_t}$ , with training sample  $\mathbf{s}, \mathbf{x}$  of *his* and test sample  $\mathbf{s}$  of *she*. Green rectangle is  $h_\theta$  as defined in Equation 7b. *Training*: Method  $\alpha$  multiplies writer style  $\bar{\mathbf{w}}$  with each character string matrix  $\mathbf{C}_{c_t}$ . Method  $\beta$  restore temporal dependencies of segment-level writer-character-DSDs ( $\mathbf{w}_h, \mathbf{w}_i, \mathbf{w}_s$ ) via an LSTM, which produces higher-quality results that are preferred by users (Sec. 4). Target test image is in red. *Runtime*: Both prediction model Method  $\alpha$  and  $\beta$  are combined to synthesize a new sample given contents within the reference sample.

via  $f_\theta^{\text{enc}}$ , provide writer-character-DSDs, e.g., *his* contains *s* and *h* in *she* and so we can estimate  $\mathbf{w}_s$  and  $\mathbf{w}_h$ . As these characters could come from any place in the reference samples,  $h_\theta$  restores the missing sequence dependencies.

### 3.1 Training losses

We defer full architecture details for our supplemental material, and here explain our losses. We begin with a point location loss  $\mathcal{L}^{\text{loc}}$  on predicted shifts in  $x, y$  coordinates,  $(\Delta x, \Delta y)$ . As we employ mixture density networks as a final prediction layer in  $f_\theta^{\text{dec}}$ , we try to maximize the probability for the target shifts  $(\Delta x^*, \Delta y^*)$  as explained by Graves et al. [17]:

$$\mathcal{L}^{\text{loc}} = - \sum_t \log \left( \sum_j \pi_t^j \mathcal{N}(\Delta x_t^*, \Delta y_t^* | \mu_{x_t}^j, \mu_{y_t}^j, \sigma_{x_t}^j, \sigma_{y_t}^j, \rho_t^j) \right).$$

Further, we consider ‘end of sequence’ flags *eos* and ‘end of character’ flags *eoc* by computing binary cross-entropy losses  $\mathcal{L}^{\text{eos}}, \mathcal{L}^{\text{eoc}}$  for each.

Next, we consider consistency in predicting writer-DSD  $\mathbf{w}$  from different writer-character-DSDs  $\mathbf{w}_{c_t}$ . We penalize a loss  $\mathcal{L}^{\mathbf{w}}$  that minimizes the variance in  $\mathbf{w}_t$  in Equation 6:

$$\mathcal{L}^{\mathbf{w}} = \sum_t (\bar{\mathbf{w}} - \mathbf{w}_t)^2 \quad (8)$$

Further, we penalize the reconstruction of each writer-character-DSD. We compare the writer-character-DSD retrieved by  $f_\theta^{\text{enc}}$  from inference samples as  $\mathbf{w}_{c_t}$  to their reconstructions  $(\mathbf{w}_{c_t}^\alpha, \mathbf{w}_{c_t}^\beta)$  via generation Methods  $\alpha$  and  $\beta$ :

$$\mathcal{L}_{A \in (\alpha, \beta)}^{\mathbf{w}_{c_t}} = \sum_t (\mathbf{w}_{c_t} - \mathbf{w}_{c_t}^A)^2 \quad (9)$$



When  $t = 1$ ,  $\mathcal{L}_\beta^{\mathbf{w}_{c_1}} = (\mathbf{w}_{c_1} - h_\theta(\mathbf{w}_{c_1}))^2$ . As such, minimizing this loss prevents  $h_\theta$  in generation Method  $\beta$  from diluting the style representation  $\mathbf{w}_{c_1}$  generated by  $f_\theta^{\text{enc}}$  because  $h_\theta$  is induced to output  $\mathbf{w}_{c_1}$ .

Each loss can be computed for three types of writer-character-DSD  $\mathbf{w}_{c_t}$ : those predicted by  $f_\theta^{\text{enc}}$ , Method  $\alpha$ , and Method  $\beta$ . These losses can also be computed at character, word, and sentence levels, e.g., for words:

$$\mathcal{L}_{\text{word}} = \sum_{A \in (f_\theta^{\text{enc}}, \alpha, \beta)} \left( \mathcal{L}_A^{\text{loc}} + \mathcal{L}_A^{\text{eos}} + \mathcal{L}_A^{\text{eoc}} + \mathcal{L}_A^{\mathbf{w}} + \mathcal{L}_A^{\mathbf{w}_{c_t}} \right). \quad (10)$$

Thus, the total loss is:  $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{char}} + \mathcal{L}_{\text{word}} + \mathcal{L}_{\text{sentence}}$ .

$\mathcal{L}_{f_\theta^{\text{enc}}}^{\mathbf{w}_{c_t}} = 0$  by construction from Equation 9; we include it here for completeness.

Sentence-level losses help to make the model predict spacing between words. While our model could train just with character- and word-level losses, this would cause a problem if we ask the model to generate a sentence from a reference sample of a single word. Training with  $\mathcal{L}_{\text{sentence}}$  lets our model predict how a writer would space words based on their writer-DSD  $\mathbf{w}$ .

*Implicit  $\mathbf{C}$  inverse constraint.* Finally, we discuss how  $\mathcal{L}^{\mathbf{w}_{c_t}}$  at the character level implicitly constrains character-DSD  $\mathbf{C}$  to be invertible. If we consider a single character sample, then mean  $\bar{\mathbf{w}}$  in Equation 6 is equal to  $\mathbf{C}_{c_1}^{-1}\mathbf{w}_{c_1}$ . In this case, as  $\mathbf{w}_{c_t}^\alpha = \mathbf{C}_{c_t}\bar{\mathbf{w}}$  (Eq. 7a),  $\mathcal{L}_\alpha^{\mathbf{w}_{c_t}}$  becomes:

$$\mathcal{L}_\alpha^{\mathbf{w}_{c_t}} = (\mathbf{w}_{c_1} - \mathbf{C}_{c_1}\mathbf{C}_{c_1}^{-1}\mathbf{w}_{c_1})^2 \quad (11)$$

This value becomes nonzero when  $\mathbf{C}$  is singular ( $\mathbf{C}\mathbf{C}^{-1} \neq \mathbf{I}$ ), and so our model avoids non-invertible  $\mathbf{C}$ s.

*Training through inverses.* As we train our network end-to-end, our model must backpropagate through  $\mathbf{C}_{c_t}$  and  $\mathbf{C}_{c_t}^{-1}$ . As derivative of matrix inverses can be obtained with  $\frac{d\mathbf{C}^{-1}}{dx} = -\mathbf{C}^{-1}\frac{d\mathbf{C}}{dx}\mathbf{C}^{-1}$ , our model can train.

## 4 Experiments

*Dataset.* Our new dataset—BRUSH—provides characteristics that other online English handwriting datasets do not, including the typical online English handwriting dataset IAM [38]. First, we explicitly display a baseline in every drawing box during data collection. This enables us to create handwriting samples whose initial action is the  $x, y$  shift from the baseline to the starting point. This additional information might also help improve performance in recognition tasks.

Second, our 170 individuals wrote 488 words *in common* across 192 sentences. This helps to evaluate handwriting models and observe whether  $\mathbf{w}$  and  $\mathbf{C}$  are decoupled: given a sample that failed to generate, we can compare the generated results of the same word across writers. If writer A failed but B succeeded, then it is likely that the problem is not with  $\mathbf{C}$  representations but with either  $\mathbf{w}$  or

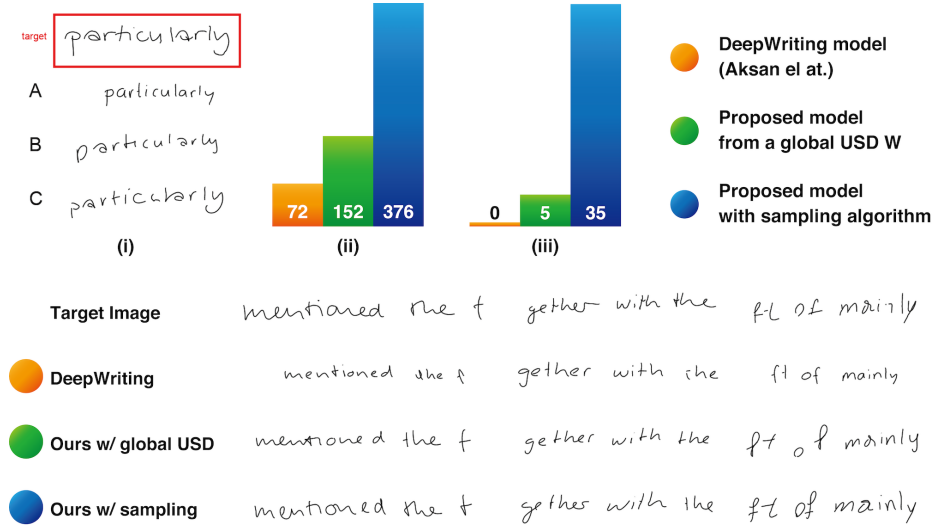


Fig. 4: Comparison of our proposed model vs. the state-of-the-art model [2]. *Top*: (i) Example writing similarity ordering task assigned to MTurk workers. (ii) Counts of most similar results with the target image. (iii) Sample-level vote. *Bottom*: Three examples of task orderings; see supplemental for all 40. The model of Aksan et al. [2] typically over-smooths the style and loses key details.

$\mathbf{w}_{c_t}$ . If both A and B failed to draw the word but succeeding in generating other words, it is likely that **C** or  $\mathbf{w}_{c_t}$  representations are to blame. We provide further details about our dataset and collection process in our supplemental material.

Third, for DeepWriting [2] comparisons, we use their training and test splits on IAM that mix writer identities—i.e., in training, we see some data from every writer. For all other experiments, we use our dataset, where we split between writers—our 20 test writers have never been seen writing *anything* in training.

*Invertibility of C.* To compute  $\mathbf{w}$  in Equation 4, we must invert the character-DSD **C**. Our network is designed to make **C** invertible as training proceeds by penalizing a reconstruction loss for  $\mathbf{w}_{c_t}$  and  $\mathbf{C}_{c_t} \mathbf{C}_{c_t}^{-1} \mathbf{w}_{c_t}$  (Sec. 3.1). To test its success, we compute **C**s from our model for all single characters (86 characters) and character pairs ( $86^2 = 7,396$  cases), and found **C** to have full rank in each case. Next, we test all possible 3-character-strings ( $86^3 = 636,056$  cases). Here, there were 37 rare cases with non-invertible **C**s, such as *1Zb* and *6ak*. In these cases, we can still extract two candidate  $\mathbf{w}$  from the first two characters (e.g., *1* and *1Z* in the *1Zb* sample) to complete generation tasks.

*Qualitative evaluation with users.* We use Amazon Mechanical Turk to asked 25 participants to rank generated handwriting similarity to a target handwriting (Fig. 4 (i)). We randomly selected 40 sentence-level target handwriting samples

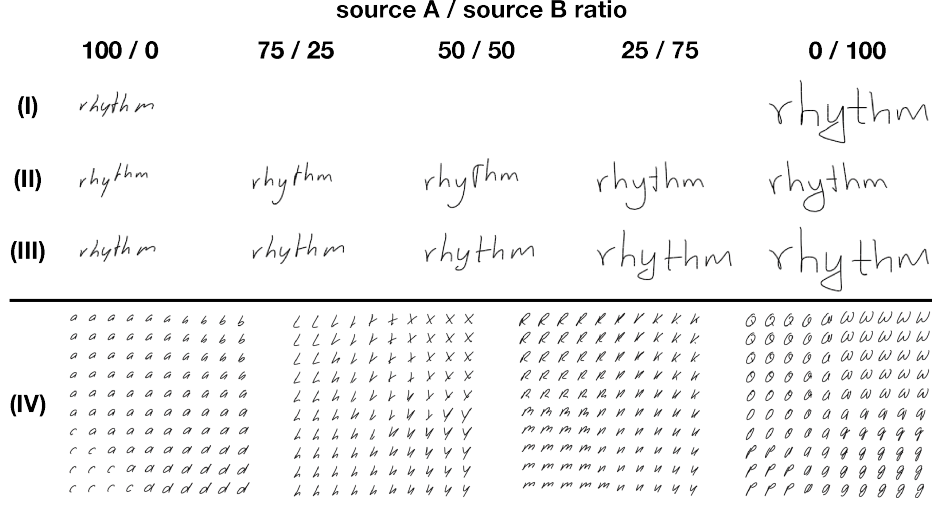


Fig. 5: Interpolation at different levels. (I) Original samples by two writers. (II) At the writer-DSD  $\mathbf{w}$  level. (III) At the writer-character-DSD  $\mathbf{w}_{c_t}$  level. (IV) At C level. *Left to right*: Characters used are *abcd*, *Lxhy*, *Rkmy*, *QWPg*.

from the validation set of IAM dataset [38]. Each participant saw randomly-shuffled samples; in total, 600 assessments were made. We compared the abilities of three models to generate the same handwriting style without seeing the actual target sample. We compare to the state-of-the-art DeepWriting model [2], which uses a sample from the same writer (but of a different character sequence) for style inference. We test both Methods  $\alpha$  and  $\beta$  from our model. Method  $\alpha$  uses the same sample to predict  $\mathbf{w}$  and to generate a new sample. Method  $\beta$  randomly samples 10 sentence-level drawing by the target writer and creates a sample with the algorithm discussed in Sec. 3. DeepWriting cannot take advantage of any additional character samples at inference time because it estimates only a single character-independent style vector.

Figure 4 (ii) displays how often each model was chosen as the most similar to the target handwriting; our model with sampling algorithm was selected  $5.22\times$  as often as Aksan et al.’s model. Figure 4 (iii) displays which model was preferred across the 40 cases: of the 15 assessments per case, we count the number of times each model was the most popular. We show all cases in supplemental material.

*Interpolation of  $\mathbf{w}$ ,  $\mathbf{w}_{c_t}$ , and C.* Figure 5 demonstrates that our method can interpolate (II) at the writer-DSD  $\mathbf{w}$  level, (III) at the writer-character-DSD  $\mathbf{w}_{c_t}$  level, and (IV) at the character-DSD C level. Given two samples of the same word by two writers  $\mathbf{x}^A$  and  $\mathbf{x}^B$ , we first extract writer-character-DSDs from each sample (e.g.,  $\mathbf{w}_{rhy}^A$ ,  $\mathbf{w}_{rhythm}^B$ ), then we derive writer-DSDs  $\mathbf{w}^A$  and  $\mathbf{w}^B$  as in Sec. 3. To interpolate by  $\gamma$  between two writers, we compute the weighted average  $\mathbf{w}^C = \gamma\mathbf{w}^A + (1 - \gamma)\mathbf{w}^B$ . Finally, we reconstruct writer-character-DSDs from

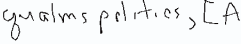
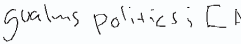

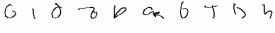
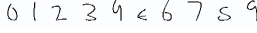
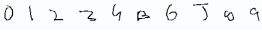
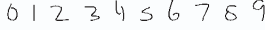
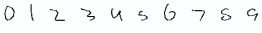
	Writer A	Writer B
Source for $\mathbf{W}$		
$\mathbf{C}$ from 1 sample		
$\mathbf{C}$ from 10 samples		
$\mathbf{C}$ from 100 samples		

Fig. 6: Predicting  $\mathbf{C}$  from new character samples, given a version of our model that is not trained on numbers. As we increase the number of samples used to estimate  $\mathbf{C}$ , the better the stylistic differences are preserved when multiplying with  $\mathbf{w}$ s from different writers A and B. *Note:* neither writers provided numeral samples; by our construction, samples can come from any writer.

$\overline{\mathbf{w}}^C$  (e.g.,  $\mathbf{w}_{rhy}^C = \mathbf{C}_{rhy} \overline{\mathbf{w}}^C$ ) and feed this into  $f_{\theta}^{dec}$  to generate a new sample. For (III), we simply interpolate at the sampled character-level (e.g.,  $\mathbf{w}_{rhy}^A$  and  $\mathbf{w}_{rhy}^B$ ). For (IV), we bilinearly interpolate four character-DSDs  $\mathbf{C}_{c_i}$  placed at the corners of each image:  $\overline{\mathbf{C}} = (r_A \times \mathbf{C}_A + r_B \times \mathbf{C}_B + r_C \times \mathbf{C}_C + r_D \times \mathbf{C}_D)$ , where all  $r$  sum to 1. From  $\overline{\mathbf{C}}$ , we compute a writer-character-DSD as  $\mathbf{w}_{\bar{c}} = \overline{\mathbf{C}}\mathbf{W}$  and synthesize a new sample. In each case (II-IV), our representations are smooth.

*Synthesis of new characters.* Our approach allows us to generate handwriting for new characters from a few samples from any writer. Let us assume that writer  $A$  produces a new character sample  $\beta$  that is not in our dataset. To make  $\beta$  available for generation in other writer styles, we need to recover the character-DSD  $\mathbf{C}_{\beta}$  that represents the shape of the character  $\beta$ . Given  $\mathbf{x}$  for newly drawn character  $\beta$ , encoder  $f_{\theta}^{enc}$  first extracts the writer-character-DSD  $\mathbf{w}_{\beta}$ . Assuming that writer  $A$  provided other non- $\beta$  samples in our dataset, we can compute multiple writer-DSD  $\mathbf{w}$  for  $A$ . This lets us solve for  $\mathbf{C}_{\beta}$  using least squares methods. We form matrices  $\mathbf{Q}, \mathbf{P}_{\beta}$  where each column of  $\mathbf{Q}$  is one specific instance of  $\mathbf{w}$ , and where each column of  $\mathbf{P}_{\beta}$  is one specific instance of  $\mathbf{w}_{\beta}$ . Then, we minimize the sum of the squared error, which is the Frobenius norm  $\|\mathbf{C}_{\beta}\mathbf{Q} - \mathbf{P}_{\beta}\|_F^2$ , e.g., via  $\mathbf{C}_{\beta} = \mathbf{P}_{\beta}\mathbf{Q}^+$ .

As architected (and detailed in supplemental),  $g_{\theta}$  actually has two parts: an LSTM encoder  $g_{\theta}^{\text{LSTM}}$  that generates a  $256 \times 1$  character representation vector  $\mathbf{c}_{c_t}^{\text{raw}}$  for a substring  $c_t$ , and a fully-connected layer  $g_{\theta}^{\text{FC2}}$  that expands  $\mathbf{c}_{c_t}^{\text{raw}}$  and reshapes it into a  $256 \times 256$  matrix  $\mathbf{C}_{c_t} = g_{\theta}^{\text{FC2}}(\mathbf{c}_{c_t}^{\text{raw}})$ . Further, as the output of an LSTM, we know that  $\mathbf{c}_{c_t}^{\text{raw}}$  should be constrained to values  $[-1, +1]$ . Thus, for this architecture, we directly optimize the (smaller set of) parameters of the latent vector  $\mathbf{c}_{c_t}^{\text{raw}}$  to create  $\mathbf{C}_{c_t}$  given the pre-trained fully-connected layer weights, using a constrained non-linear optimization algorithm (L-BFGS-B) via the objective  $f(\mathbf{c}_{c_t}^{\text{raw}}) = \|\mathbf{P}_{\beta} - g_{\theta}^{\text{FC2}}(\mathbf{c}_{c_t}^{\text{raw}})\mathbf{Q}\|_F^2$ .

To examine this capability of our approach, we retrained our model with a modified dataset that *excluded* numbers. In Figure 6, we see generation using our

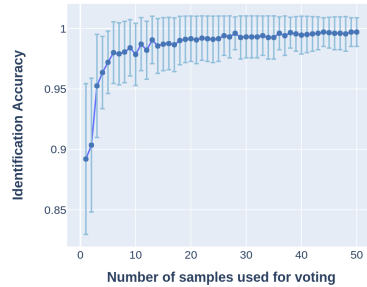
estimate of new  $C$ s from different sample counts. We can generate numerals in the style of a particular writer even though they never drew them, using relatively few drawing samples of the new characters from *different* writers.

*Writer recognition task.* Writer recognition systems try to assign test samples (e.g., a page of handwriting) to a particular writer given an existing database. Many methods use codebook approaches [8,11,22,7] to catalogue characteristic patterns such as graphemes, stroke junctions, and key-points from offline handwriting images and compare them to test samples. Zhang et al. [45] extend this idea to online handwriting, and Adak et al. study idiosyncratic character style per person and extract characteristic patches to identify the writer [1].

To examine how well our model might represent the latent distribution of handwriting styles, we perform a writer recognition task on our trained model on the randomly-selected 20-writer hold out set from our dataset. First, we compute 20 writer DSDs  $\bar{\mathbf{w}}_i^{writer}$  from 10 sentence-level samples—this is our offline ‘codebook’ representing the style of each writer. Then, for testing, we sample from 1–50 new word-level stroke sequences per writer (using words with at least 5 characters), and calculate the corresponding writer DSDs ( $N = 1,000$  in total). With the vector  $L$  of true writer labels, we compute prediction accuracy:

$$A = \frac{1}{N} \sum_{i=1}^N I(L_i, \arg \min_j (\bar{\mathbf{w}}_i^{word} - \bar{\mathbf{w}}_j^{writer})^2), I(x, y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

We repeated the random sampling of 1–50 words over 100 trials and compute mean accuracy and standard error. When multiple test samples are provided, we predict writer identity for each word and average their predictions. Random accuracy performance is 5%. Our test prediction accuracy rises from  $89.20\% \pm 6.23$  for one word sample, to  $97.85\% \pm 2.57$  for ten word samples, to  $99.70\% \pm 1.18$  for 50 word samples. Increasing



the number of test samples per writer increases accuracy because some words may not be as characteristic as others (e.g., ‘tick’ vs. ‘anon’). Overall, while our model was not trained to perform this classification task, we can still achieved promising accuracy results from few samples—this is an indication that our latent space is usefully descriptive.

*Additional experiments.* In our supplemental material, along with more architecture, model training procedure, and sampling algorithm details, we also: 1) compare to two style extraction pipelines, a stacked FC+ReLU layers and AdaIN, and find our approach more capable; 2) demonstrate the importance of learning style and content of character-DSD  $\mathbf{C}$  by comparing with a randomly-initialized version; 3) ablate parts of our loss function, and illustrate key components; 4) experimentally show that our model is more efficient than DeepWriting by comparing generation given the same number of model parameters.

## 5 Discussion

While users preferred our model in our study, it still sometimes fails to generate readable letters or join cursive letters. One issue here is the underlying inconsistency in human writers, which we only partially capture in our data and represent in our model (e.g., cursive inconsistency). Another issue is collecting high-quality data with digital pens in a crowdsourced setting, which can still be a challenge and requires careful cleaning (see supplemental for more details).

*Decoupling additional styles.* Our model could potentially scale to more styles. For instance, we might create an age matrix  $\mathbf{A}$  from a numerical age value  $a$  as  $\mathbf{C}$  is constructed from  $c_t$ , and extract character-independent age-independent style descriptor as  $\mathbf{w}^* = \mathbf{A}^{-1}\mathbf{C}_{c_t}^{-1}\mathbf{w}_{c_t}$ . Introducing a new age operator  $\mathbf{A}$  invites our model to find latent-style similarities across different age categories (e.g., between a child and a mature writer). Changing the age value and thus  $\mathbf{A}$  may predict how a child’s handwriting changes as s/he becomes older. However, training multiple additional factors in this way is likely to be challenging.

*Alternatives to linear  $\mathbf{C}$  multiplication operator.* Our model can generate new characters by approximating a new  $\mathbf{C}$  matrix from few pairs of  $\mathbf{w}$  and  $\mathbf{w}_{c_t}$  thanks to their linear relationship. However, one might consider replacing our matrix multiplication ‘operator’ on  $\mathbf{C}$  with parametrized nonlinear function approximators, such as autoencoders. Multiplication by  $\mathbf{C}^{-1}$  would become an encoder, with multiplication by  $\mathbf{C}$  being a decoder; in this way,  $g_\theta$  would be tasked with predicting encoder weights given some predefined architecture. Here, consistency with  $\mathbf{w}$  must still be retained. We leave this for future work.

## 6 Conclusion

We introduce an approach to online handwriting stroke representation via the Decoupled Style Descriptor (DSD) model. DSD succeeds in generating drawing samples which are preferred more often in a user study than the state-of-the-art model. Further, we demonstrate the capabilities of our model in interpolating samples at different representation levels, recovering representations for new characters, and achieving a high writer-identification accuracy, despite not being trained explicitly to perform these tasks. Online handwriting synthesis is still challenging, particularly when we infer a stylistic representation from few numbers of samples and try to generate new samples. However, we show that decoupling style factors has potential, and believe it could also apply to style-related tasks like transfer and interpolation in other sequential data domains, such as in speech synthesis, dance movement prediction, and musical understanding.

*Acknowledgements.* This work was supported by the Sloan Foundation and the National Science Foundation under award number IIS-1652561. We thank Kwang In Kim for fruitful discussions and for being our matrix authority. We thank Naveen Srinivasan and Purvi Goel for the ECCV deadline snack delivery service. Finally, we thank all anonymous writers who contributed to our dataset.

## References

1. Adak, C., Chaudhuri, B.B., Lin, C.T., Blumenstein, M.: Intra-variable handwriting inspection reinforced with idiosyncrasy analysis (2019) **13**
2. Aksan, E., Pece, F., Hilliges, O.: DeepWriting: Making Digital Ink Editable via Deep Generative Modeling. In: SIGCHI Conference on Human Factors in Computing Systems. CHI '18, ACM, New York, NY, USA (2018) **1, 2, 3, 4, 10, 11, 19, 23, 30**
3. Alonso, E., Moysset, B., Messina, R.O.: Adversarial generation of handwritten text images conditioned on sequences. ArXiv [abs/1903.00277](https://arxiv.org/abs/1903.00277) (2019) **3**
4. Azadi, S., Fisher, M., Kim, V.G., Wang, Z., Shechtman, E., Darrell, T.: Multi-content gan for few-shot font style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7564–7573 (2018) **3**
5. Balashova, E., Bermanno, A.H., Kim, V.G., DiVerdi, S., Hertzmann, A., Funkhouser, T.: Learning a stroke-based representation for fonts. Computer Graphics Forum **38**(1), 429–442 (2019). <https://doi.org/10.1111/cgf.13540>, <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13540> **2**
6. Baluja, S.: Learning typographic style: from discrimination to synthesis. Machine Vision and Applications **28**(5-6), 551–568 (2017) **3**
7. Bennour, A., Djeddi, C., Gattal, A., Siddiqi, I., Mekhaznia, T.: Handwriting based writer recognition using implicit shape codebook. Forensic science international **301**, 91–100 (2019) **13**
8. Bensefia, A., Nosary, A., Paquet, T., Heutte, L.: Writer identification by writer's invariants. In: Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition. pp. 274–279. IEEE (2002) **13**
9. Berio, D., Akten, M., Leymarie, F.F., Grierson, M., Plamondon, R.: Calligraphic stylisation learning with a physiologically plausible model of movement and recurrent neural networks. In: Proceedings of the 4th International Conference on Movement Computing. pp. 1–8 (2017) **3**
10. Bishop, C.M.: Mixture density networks (1994) **3, 5, 29**
11. Bulacu, M., Schomaker, L.: Text-independent writer identification and verification using textural and allographic features. IEEE transactions on pattern analysis and machine intelligence **29**(4), 701–717 (2007) **13**
12. Campbell, N.D., Kautz, J.: Learning a manifold of fonts. ACM Transactions on Graphics (TOG) **33**(4), 1–11 (2014) **2**
13. Carter, S., Ha, D., Johnson, I., Olah, C.: Experiments in handwriting with a neural network. Distill (2016). <https://doi.org/10.23915/distill.00004>, <http://distill.pub/2016/handwriting> **3**
14. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (Oct 2014). <https://doi.org/10.3115/v1/D14-1179>, <https://www.aclweb.org/anthology/D14-1179> **4, 5**
15. Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A.C., Bengio, Y.: A recurrent latent variable model for sequential data. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 2980–2988. Curran Associates, Inc. (2015), <http://papers.nips.cc/paper/5653-a-recurrent-latent-variable-model-for-sequential-data.pdf> **4**

16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014) [3](#)
17. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013) [1](#), [3](#), [8](#)
18. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning. p. 369–376. ICML '06, Association for Computing Machinery, New York, NY, USA (2006). <https://doi.org/10.1145/1143844.1143891>, <https://doi.org/10.1145/1143844.1143891> [30](#)
19. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE transactions on pattern analysis and machine intelligence **31**(5), 855–868 (2008) [30](#)
20. Ha, D., Dai, A., Le, Q.V.: Hypernetworks. arXiv preprint arXiv:1609.09106 (2016) [3](#)
21. Haines, T.S.F., Mac Aodha, O., Brostow, G.J.: My text in your handwriting. ACM Trans. Graph. **35**(3) (May 2016). <https://doi.org/10.1145/2886099>, <https://doi.org/10.1145/2886099> [3](#), [4](#)
22. He, S., Wiering, M., Schomaker, L.: Junction detection in handwritten documents and its application to writer identification. Pattern Recognition **48**(12), 4036–4048 (2015) [13](#)
23. Heigold, G., Moreno, I., Bengio, S., Shazeer, N.: End-to-end text-dependent speaker verification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5115–5119. IEEE (2016) [4](#)
24. Hsu, W.N., Glass, J.: Scalable factorized hierarchical variational autoencoder training. arXiv preprint arXiv:1804.03201 (2018) [4](#)
25. Hsu, W.N., Zhang, Y., Glass, J.: Unsupervised learning of disentangled and interpretable representations from sequential data. In: Advances in Neural Information Processing Systems (2017) [4](#)
26. Hu, C., Hersch, R.D.: Parameterizable fonts based on shape components. IEEE Computer Graphics and Applications **21**(3), 70–85 (2001) [2](#)
27. Hu, Q., Szabó, A., Portenier, T., Favaro, P., Zwicker, M.: Disentangling factors of variation by mixing them. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) [3](#)
28. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017) [2](#), [19](#)
29. Jaeger, S., Manke, S., Reichert, J., Waibel, A.: Online handwriting recognition: the npen++ recognizer. International Journal on Document Analysis and Recognition **3**(3), 169–180 (2001) [30](#)
30. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. In: Advances in neural information processing systems. pp. 667–675 (2016) [3](#), [6](#)
31. Keysers, D., Deselaers, T., Rowley, H.A., Wang, L.L., Carbune, V.: Multi-language online handwriting recognition. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1180–1194 (2016) [30](#)
32. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014) [32](#)
33. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) [4](#)



34. Kotovenko, D., Sanakoyeu, A., Lang, S., Ommer, B.: Content and style disentanglement for artistic style transfer. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4422–4431 (2019) [2](#)
35. Lahiri, S.: Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In: Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 96–105. Association for Computational Linguistics, Gothenburg, Sweden (April 2014), <http://www.aclweb.org/anthology/E14-3011> [33](#)
36. Lian, Z., Zhao, B., Chen, X., Xiao, J.: Easyfont: a style learning-based system to easily build your large-scale handwriting fonts. *ACM Transactions on Graphics (TOG)* **38**(1), 1–18 (2018) [3](#)
37. Lopes, R.G., Ha, D., Eck, D., Shlens, J.: A learned representation for scalable vector graphics. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) [2](#)
38. Marti, U.V., Bunke, H.: The iam-database: an english sentence database for off-line handwriting recognition. *International Journal on Document Analysis and Recognition* **5**(1), 39–46 (2002) [2](#), [5](#), [9](#), [11](#)
39. Oord, A.v.d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016) [4](#)
40. Qian, K., Zhang, Y., Chang, S., Yang, X., Hasegawa-Johnson, M.: Autovc: Zero-shot voice style transfer with only autoencoder loss. In: International Conference on Machine Learning. pp. 5210–5219 (2019) [2](#)
41. Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al.: Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4779–4783. IEEE (2018) [4](#)
42. Stanley, K.O., D’Ambrosio, D.B., Gauci, J.: A hypercube-based encoding for evolving large-scale neural networks. *Artificial life* **15**(2), 185–212 (2009) [3](#)
43. Suveeranont, R., Igarashi, T.: Example-based automatic font generation. In: International Symposium on Smart Graphics. pp. 127–138. Springer (2010) [3](#)
44. Wang, H., Liang, X., Zhang, H., Yeung, D.Y., Xing, E.P.: Zm-net: Real-time zero-shot image manipulation network. *arXiv preprint arXiv:1703.07255* (2017) [3](#), [6](#)
45. Zhang, X.Y., Xie, G.S., Liu, C.L., Bengio, Y.: End-to-end online writer identification with recurrent neural network. *IEEE Transactions on Human-Machine Systems* **47**(2), 285–292 (2016) [13](#)
46. Zhang, X.Y., Yin, F., Zhang, Y.M., Liu, C.L., Bengio, Y.: Drawing and recognizing chinese characters with recurrent neural network. *IEEE transactions on pattern analysis and machine intelligence* **40**(4), 849–862 (2017) [3](#)
47. Zongker, D.E., Wade, G., Salesin, D.H.: Example-based hinting of true type fonts. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 411–416 (2000) [2](#)

## Appendices for Generating Handwriting via Decoupled Style Descriptors

A	Table of Variables . . . . .	18
B	Comparison with Style Transfer Baselines . . . . .	19
C	Investigating the $\mathbf{C}$ -matrix . . . . .	19
D	Network Capacity . . . . .	23
E	Further Generated Comparisons . . . . .	24
F	Sampling Algorithm for Writer-Character-DSD $\mathbf{w}_{c_t}$ . . . . .	27
G	Sequence Decoder $f_{\theta}^{\text{dec}}$ . . . . .	28
H	Character Encoder Function $g_{\phi}$ . . . . .	29
I	Segmentation Network $k_{\theta}$ . . . . .	30
J	Detailed Training Procedure . . . . .	31
K	Dataset Specification and Collection Methodology . . . . .	32

### A Table of Variables

We include a brief table of the key variables used throughout the main manuscript and in this supplemental manuscript (Table 2).

Table 2: Brief explanation of key variables used throughout these manuscripts.

	Name	Shape Note
$\mathbf{x}$	Input data	$(N, 3)$ A handwriting sample; a time sequence of 2D points.
$\mathbf{x}^*$	Encoded input	$(N, 256)$ A raw output from $f_{\theta}^{\text{enc}}(\mathbf{x})$ .
$s$	Sentence	$(M)$ A string label for $\mathbf{x}$ (e.g., <i>hello</i> ).
$c_t$	Substring	$(t)$ A substring of $s$ (e.g., <i>he</i> ).
$\mathbf{c}_t$	Character vector	$(87 \times 1)$ A one-hot vector denoting the $t$ -th character in $s$ .
$\mathbf{c}_t^{\text{raw}}$	Encoded character	$(256 \times 1)$ An output from $g_{\theta}^{\text{FC1}}(\mathbf{c}_t)$ . Input for $g_{\theta}^{\text{LSTM}}$ .
$\mathbf{c}_{c_t}^{\text{raw}}$	Encoded substring	$(256 \times 1)$ An output from $g_{\theta}^{\text{LSTM}}(\mathbf{c}_t^{\text{raw}})$ .
$\mathbf{w}$	Writer-DSD	$(256 \times 1)$ Content-independent handwriting style for a writer $A$ .
$\mathbf{C}_{c_t}$	Character-DSD	$(256 \times 256)$ An encoded character matrix for a substring $c_t$ .
$\mathbf{w}_{c_t}$	Writer-Character-DSD	$(256 \times 1)$ An encoded drawing representation for $c_t$ , extracted from $\mathbf{x}^*$ .
$f_{\theta}^{\text{enc}}$	Sequence encoder	Outputs a list of Writer-Character-DSDs $\mathbf{w}_{c_t}$ from an input drawing $\mathbf{x}$ .
$f_{\theta}^{\text{dec}}$	Sequence decoder	Outputs a drawing $\mathbf{x}$ from a list of $\mathbf{w}_{c_t}$ .
$g_{\theta}$	Character encoder	Outputs a character matrix $\mathbf{C}_{c_t}$ . Simplified function used as shorthand.
$g_{\theta}^{\text{FC1}}$		Outputs a vector $\mathbf{c}_t^{\text{raw}}$ from a vector $\mathbf{c}_t$ .
$g_{\theta}^{\text{LSTM}}$		Outputs a vector $\mathbf{c}_{c_t}^{\text{raw}}$ from a list $[\mathbf{c}_1^{\text{raw}}, \dots, \mathbf{c}_t^{\text{raw}}]$ .
$g_{\theta}^{\text{FC2}}$		Outputs a Character-DSD $\mathbf{C}_{c_t}$ from a vector $\mathbf{c}_{c_t}^{\text{raw}}$ .
$h_{\theta}$	Temporal encoder	LSTM to restore dependencies between Writer-Character DSDs $\mathbf{w}_{c_t}$ .
$k_{\theta}$	Segmentation function	Segments a handwriting sample $\mathbf{x}$ into characters.

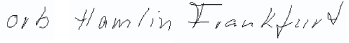



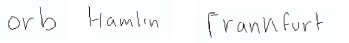


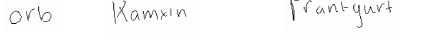
	Writer A	Writer B
Target		
Ours (a)		
Method A		
Method B		

Fig. 7: Qualitative evaluation of two common style-transfer techniques.

## B Comparison with Style Transfer Baselines

We evaluated our proposed model against two style transfer baselines. We define a style vector as  $\mathbf{s} = f_{\theta}^{\text{enc}}(\mathbf{x})$  and a character-content vector as  $\mathbf{c} = g_{\theta}(c_t)$ . To interweave  $\mathbf{s}$  and  $\mathbf{c}$ , we consider a new operator  $F$  where  $F(\mathbf{s}, \mathbf{c}) = z$ . Then, we feed  $z$  into our decoder function  $f_{\theta}^{\text{dec}}$  to synthesize a drawing. We examined two operators for  $F$ : A) three stacked FC+ReLU layers, and B) AdaIN layer [28].

In our method,  $f_{\theta}^{\text{enc}}(\mathbf{x})$  produces  $\mathbf{w}_{c_t}$ , which is then decoupled from the character content via our  $\mathbf{C}$  matrix operation. In Method A and B,  $f_{\theta}^{\text{enc}}(\mathbf{x})$  produces  $\mathbf{s}$ , and via  $F$  the network must implicitly represent content and writer style parts. For fairness, we keep the architectures of  $f_{\theta}^{\text{enc}}$ ,  $f_{\theta}^{\text{dec}}$ ,  $g_{\theta}$  the same as in our approach, and train each method from scratch with the same data and loss function as in our approach.

Neither Method A or B is competitive with our method or with DeepWriting [2]. While A and B can generate readable letters, A) has only one style, and B) fails to capture important character shape details leaving some illegible, and has only basic style variation like slant and size (Figure 7). This is because  $f_{\theta}^{\text{enc}}$  must represent a content-independent style for a reference sample without its content information. The DeepWriting model decouples style and content by making  $f_{\theta}^{\text{enc}}$  additionally predict the reference content via a character classification loss. Our approach does not try to decouple style and content within  $f_{\theta}^{\text{enc}}$ ; instead our model extracts style from the output of  $f_{\theta}^{\text{enc}}$  by multiplication with a content-conditioned matrix  $\mathbf{C}$ .

This simple experiment demonstrates that style-content decoupling is a difficult task. Instead of making one network (i.e.,  $f_{\theta}^{\text{enc}}$ ) responsible for filtering out content information from the style reference sample implicitly, we show empirically that our method to structurally decouple content information via  $\mathbf{C}$  matrix multiplication is more effective in the online handwriting domain.

## C Investigating the C-matrix

The  $\mathbf{C}$  matrix for a character string  $c_t$ — $\mathbf{C}_{c_t}$ —is designed to contain information about how people generally write  $c_t$ : its role is to extract character(s)-specific

information from  $\mathbf{w}_{c_t}$ . Intuitively, the relationship between  $\mathbf{C}_{c_t}$  and  $\mathbf{w}_{c_t}$  can be seen as one of a key and a key-hole. Our model tries to create a perfect fit between a key ( $\mathbf{w}_{c_t}$ ) and a key-hole ( $\mathbf{C}_{c_t}$ ), where both shapes are learned simultaneously. But what if we fix the key-hole shape ahead of time, and simply learn to fit the key? That is, what if we assign pre-defined values to substring character matrices  $\mathbf{C}$  ahead of time? This would reduce the number of model parameters, speed up training and inference, and allow us to store  $\mathbf{C}$  in memory as a look-up table rather than predict its values.

One issue with fixing the  $\mathbf{C}$  matrix is the exponential growth in the number of possible strings  $c_t$  as we allow longer words. Thus, for this analysis, we will initialize  $\mathbf{C}$  for single- and two-character substrings only, which have a tractable number of variations in our Latin alphabet (Sec. K). For example, instead of  $\mathbf{C}_{hello}$  for a word ‘hello’, we consider its five constituent single- and two-character substrings  $\mathbf{C}_h$ ,  $\mathbf{C}_{he}$ ,  $\mathbf{C}_{el}$ ,  $\mathbf{C}_{il}$ ,  $\mathbf{C}_{lo}$ . Consequently, we modified the training data format by segmenting every sentence into two-character pairs.

We consider three scenarios (Figure 8):

**Fixed random single- and two-character  $\mathbf{C}$**  In principle, each substring that  $\mathbf{C}$  represents only needs to be *different* from other substrings, and so we assign a random matrix to each two-character substring.

**Fixed well-spaced single- and two-character  $\mathbf{C}$**  Two matrices  $\mathbf{C}_{sh}$  and  $\mathbf{C}_{he}$  could contain mutual information about how to write the character  $h$ , and so we try to assign fixed matrices in a way that places similar substrings close to each other in high-dimensional space.

**Learned single- and two-character  $\mathbf{C}$**  Our model trained only on single characters and two-character pairs. This trains  $g_\theta$  to predict the values of  $\mathbf{C}$ .

*Well-spaced  $\mathbf{C}$ .* If we randomly initialize  $\mathbf{C}$  (i.e.,  $I(\mathbf{C}_{sh}; \mathbf{C}_{he}) \approx 0$ ), the values of two writer-character-DSDs,  $\mathbf{w}_{he}$  and  $\mathbf{w}_{she}$ , must be significantly different from each other to output consistent  $\mathbf{w}$ , and this makes the learning task harder for the  $f_\theta^{\text{enc}}$  LSTM. Instead, to determine how to manually initialize  $\mathbf{C}$  such that they are more well spaced out, we look at the character information within  $\mathbf{w}_{c_t}$ . As we use an LSTM to encode the input drawing to obtain  $\mathbf{w}_{c_t}$ , it models long temporal dependencies. In other words, by the nature of LSTMs,  $\mathbf{w}_{c_t}$  tends to ‘remember’ more recent characters than older characters, and so we assume  $\mathbf{w}_{c_t}$  remembers the second character more than the first character. Thus, we initialize the character matrix for a two-character substring  $c_t$  as follows:

$$\mathbf{C}_{c_t} = r\mathbf{C}_{c_1} + (1.0 - r)\mathbf{C}_{c_2} \quad (13)$$

where  $\mathbf{C}_{c_1}$  and  $\mathbf{C}_{c_2}$  are randomly initialized single-character-DSDs, and we set  $r = 0.1$ . This leads to  $\mathbf{C}_{c_t}$  that have the same ending character (i.e.,  $c_2$ ) having similar representations, as shown in Figure 8b.

*Results.* Under t-SNE projections, the learned  $\mathbf{C}$  models create more meaningful  $\mathbf{C}$  representation layouts (Figure 8). Unlike the well-spaced  $\mathbf{C}$ , when we project  $\mathbf{C}$  for two-character substrings (Figure 8c), we see a few outer clusters, with

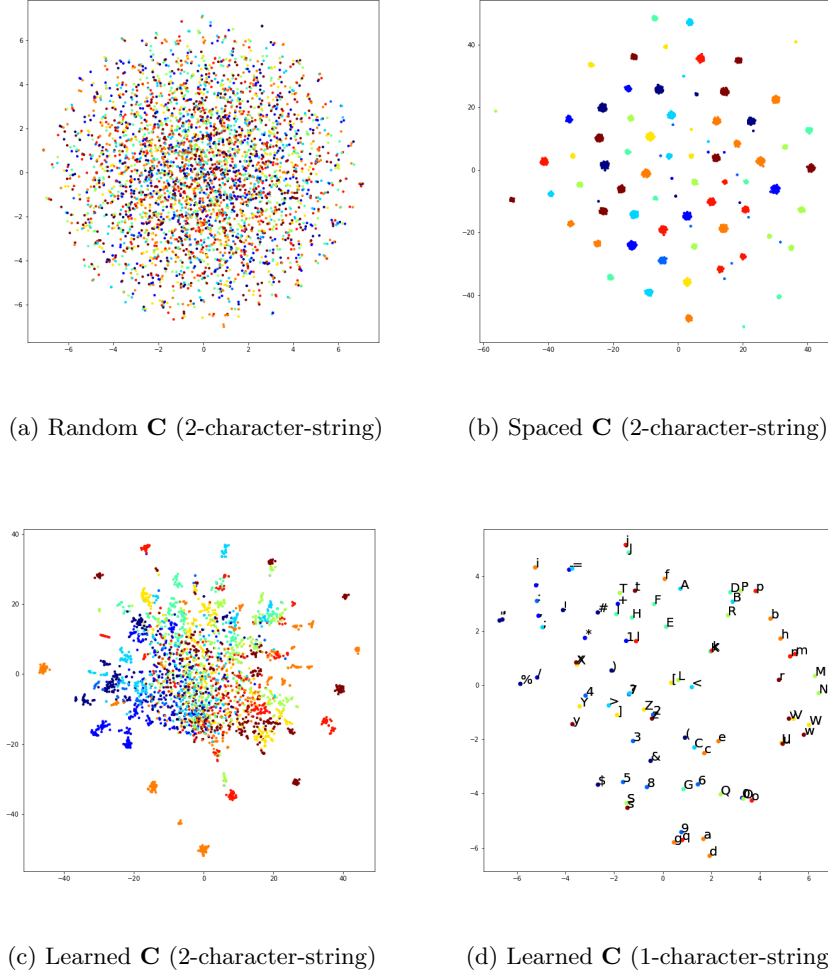


Fig. 8: t-SNE visualization of different  $\mathbf{C}$ . Each dot indicates different substring. The substrings with the same last character (e.g., ‘ab’, ‘bb’, ‘cb’) are colored same. The learned  $\mathbf{C}$  in Figure 8c are mostly concentrated in the middle. As each  $\mathbf{C}$  contains information about how to draw *two* characters, even the two  $\mathbf{C}$  with the same last character (e.g., ‘ab’ and ‘bb’) are often distant from each other, because of the different first character. By contrast, isolating the single characters within the learned  $\mathbf{C}$  (Figure 8d) shows them to be well mapped in the space: similar characters such as ‘(’, ‘c’ and ‘C’ are closely positioned.

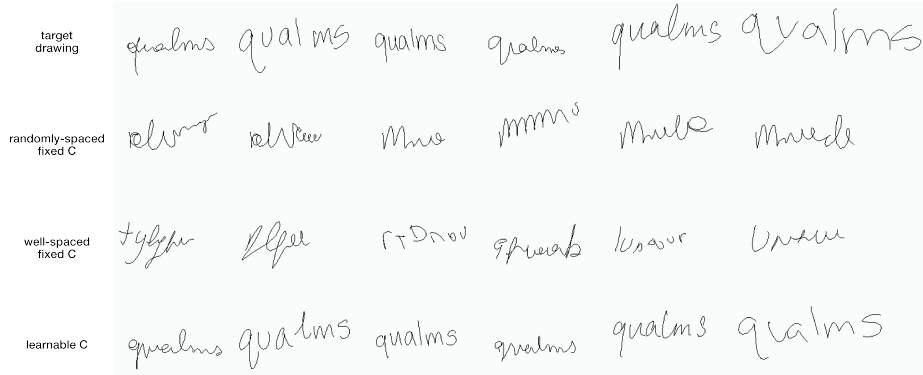


Fig. 9: Qualitative comparisons of results from different two-character **C**. When **C** are fixed through training, the models failed to synthesize recognizable letters.

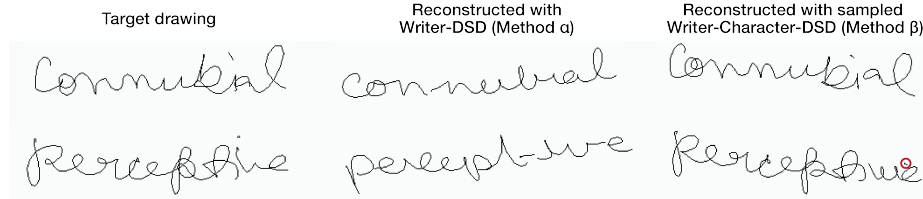


Fig. 10: Instances of missed delayed strokes by our proposed model.

a larger ‘more chaotic’ central concentration. As each dot in the projections represents **C** for two characters, these **C** cannot be easily clustered by the ending characters (e.g., considering general shapes, ‘**cb**’ is likely to have a representation closer to the one of ‘**C6**’ than ‘**fb**’, despite the common 2nd character **b**). When looking at just the single-characters within our learned **C** representations (Figure 8d), characters with similar shapes (e.g., ‘9’, ‘q’, ‘g’) are closely positioned, and this indicates a successful representation learning for **C**.

Figure 9 shows writing generation results from these different approaches. Both fixed **C** approaches fail to generate good samples.

*Limitations of two-character substrings.* One might think that using single- and two-character substrings could represent most variation in writing—how much do letters two behind the currently-written letter really affect the output? Cursive writing especially contains delayed strokes: for example, adding the dot for ‘i’ in ‘himself’ after writing ‘f’. Changing to two-character substrings removes the ability of our model to learn delayed strokes in writings. In practice, our original **C** model can struggle to correctly place delayed strokes (Fig. 10): the model must predict a negative x-axis stroke to finish a previous character, which rarely occurs in our training set. This is one area for future work to improve.

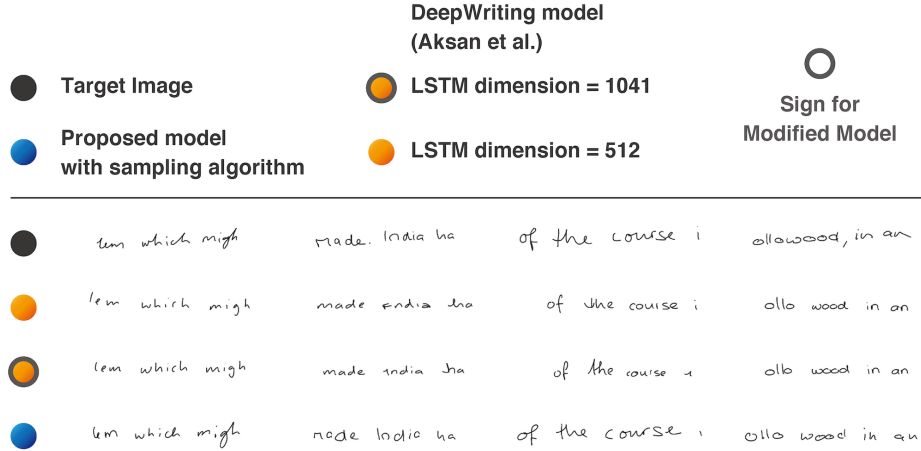


Fig. 11: To match the total number of parameters between DeepWriting and our model, we increased the LSTM dimension in DeepWriting from 512 to 1,041. There is little improvement in quality from the initial DeepWriting model of 512 LSTM dimension to 1,041. These drawings are generated with 10 sentence-level reference samples of the same writer.

## D Network Capacity

To validate our network capacity, we conducted two comparison studies with the DeepWriting model by Aksan et al. [2]. The first is to decrease the number of parameters in our DSD-based model, and the second is to increase the number of parameters in the DeepWriting model.

### A Increasing DeepWriting Model Parameters

We modified the hidden state dimension for the DeepWriting model from 512 to 1,071, and the total number of parameters subsequently increased from 7.2M to 31.3M. We show a side-by-side comparison of generated samples with DSD-256 model in Figure 11. For our 8GB VRAM GPU to accommodate this large LSTM, we decreased the batch size by half from 64 to 32, and doubled the number of learning rate decay steps from 1,000 to 2,000. However, increasing the capacity of the DeepWriting model did not improve the generated results (Figure 11).

### B Decreasing DSD Model Parameters

Is our decoupled model more efficient than the DeepWriting model [2], or simply more capacitive? With 256-dimensional latent vectors, our model has 31.33M parameters, whereas DeepWriting has 7.27M. This difference is largely in the  $g_\theta$  fully connected layer which expands  $\mathbf{c}_{c_t}^{\text{raw}}$  into  $\mathbf{C}_{c_t}$  via 16.84M parameters. As such, we reduced our latent vector dimension from 256 to 141, which leads



















	Proposed model from a global DSD W		Proposed model with sampling algorithm	
 Target Image	 DSD dimension = 141	 DSD dimension = 141		
 DeepWriting model (Aksan et al.)	 DSD dimension = 256	 DSD dimension = 256		
<hr/>				
	particularly	of the course i	earth, grant	ft of mainly
	particularly	of the course i	earth marit	ft of mainly
	particularly	of the course i	earth marit	ft of mainly
	particularly	of the course i	earth marit	ft of mainly
	particularly	of the course i	earth marit	ft of mainly
	particularly	of the course i	earth marit	ft of mainly
<hr/>				
	any tomato pla	rd of the job of the	the Connecticut	and not the one
	any tomato pla	rd of the job of the	the Connecticut	and not the one
	any tomato pla	rd of the job of the	the Connecticut	and not the one
	any tomato pla	rd of the job of the	the Connecticut	and not the one
	any tomato pla	rd of the job of the	the Connecticut	and not the one
	any tomato pla	rd of the job of the	the Connecticut	and not the one

Fig. 12: We decreased the DSD dimension in our model from 256 to 141 to match the total number of parameters to DeepWriting. As we decrease DSD dimensions, there is a slight fall in quality, particularly the examples with green dots that are generated from a single global writer-DSD  $\bar{w}$  in Method  $\alpha$ .

to a model with  $7.25M$  parameters. While we observe minor deterioration in generation quality (Figure 12), the model still creates higher-quality samples than DeepWriting. This suggests that our architecture is more efficient.

## E Further Generated Comparisons

Figure 13 and 14 show all 40 samples of drawing used for our qualitative/quantitative study on Amazon Mechanical Turk.



	Target Image	DeepWriting model (Aksan et al.)	Proposed model from a global DSD W	Proposed model with sampling algorithm
		of the course i	mentioned the f	"integrated" + she over
		of the course i	mentioned she f	integrated + ake over
		of the course i	mentioned the f	integrated + ake over
		of the course i	mentioned the f	integrated + ake over
		say that i don	he told them, "the po	series of when they saw "
		say that i don	he told them the po	series of when they saw "
		say that i don	he told them the po	series of when they saw "
		say that i don	he told them the po	series of when they saw "
		and not the one	allowood, in an	allowed val as the funds for which they
		and not the one	allo wood in an	allowed val as the funds for which they
		and not the one	allo wood in an	allowed val as the funds for which they
		and not the one	allo wood in an	allowed val as the funds for which they
		k men who	ft of mainly	on going to made India ha
		k men who	ft of mainly	on going to made India ha
		k men who	ft of mainly	on going to made India ha
		k men who	ft of mainly	on going to made India ha
		as via the tube L	tem which migh	problems an a Lower level. Co
		as via the tube L	tem which migh	problems an a Lower level Co
		as via the tube L	tem which migh	problems an a Lower level Co
		as via the tube L	tem which migh	problems an a Lower level Co

Fig. 13: The first 20 out of 40 samples used for quantitative evaluation.

●	particularly	d to see her	the Connecticut	Tomorrow he wo
●	particularly	d to see her	the connecticut	Tomorrow he wo
●	particularly	d to see her	the Connecticut	Tomorrow he wo
●	particularly	d to see her	the Connecticut	Tomorrow he wo
●	way that th	any tomato pla	nd girls aged	nd sit wit
●	way that sh	any tomato pla	nd girls aged	nd sit wit
●	way that th	any tomato pla	nd girls aged	nd sit wit
●	way that th	any tomato pla	nd girls aged	nd sit wit
●	gether with the	young offende	earth, guarit	It faced ou
●	gether with the	young offende	earth marit	It faced on
●	gether with the	young offende	earth marit	It faced on
●	gether with the	young offende	earth marit	It faced on
●	- cannot be fo	deeds follo	James ha	rd of the job of the
●	cannot be fo	deeds follo	James ha	rd of the job of the
●	cannot be fo	deeds follo	James ha	rd of the job of the
●	cannot be fo	deeds follo	James ha	rd of the job of the
●	er should be	as of off - the - c	ete bits.	hanging open
●	er should be	as of off the c	ete bits	hanging open
●	er should be	as of off the c	ete bits	hanging open
●	er should be	as of off the c	ete bits	hanging open

Fig. 14: The second 20 out of 40 samples used for quantitative evaluation.



Fig. 15: Generated image by our sampling algorithm. The black letters in the synthesis indicate that they are predicted from  $\bar{\mathbf{w}}$ , while the colored characters in reference samples are encoded and saved in the database in the form of writer-character-DSDs  $\mathbf{w}_{c_t}$  and retrieved during synthesis.

## F Sampling Algorithm for Writer-Character-DSD $\mathbf{w}_{c_t}$

When handwriting samples  $\mathbf{x}$  with corresponding character strings  $s$  are provided for inference, we can extract writer-character-DSDs  $\mathbf{w}_{c_t}$  from  $\mathbf{x}$  for substrings of  $s$ . For example, for character string *his*, we can first extract the following 3 arrays of writer-character-DSDs using  $f_{\theta}^{enc}$ :  $[\mathbf{w}_h]$ ,  $[\mathbf{w}_h, \mathbf{w}_{hi}]$ , and  $[\mathbf{w}_h, \mathbf{w}_{hi}, \mathbf{w}_{his}]$ . In addition, if the handwriting is non-cursive and each character is properly segmented, then we can also obtain 3 more ( $[\mathbf{w}_i]$ ,  $[\mathbf{w}_i, \mathbf{w}_{is}]$ , and  $[\mathbf{w}_s]$ ). However, we must ensure that the handwriting is cursive, as  $h$ ,  $i$ , and  $s$  could be connected by a single stroke. In such cases, we only extract the first 3 arrays.

We create a database  $D$  of these arrays of writer-character-DSDs with substrings as their keys, and query substrings in the target sentence  $s^*$  for generation to obtain relevant writer-character-DSDs. We also compute the mean global writer-DSD  $\bar{\mathbf{w}}$  as  $\bar{\mathbf{w}} = \frac{1}{N} \sum_{c_t} \mathbf{C}_{c_t}^{-1} \mathbf{w}_{c_t}$  where  $N$  is the number of obtained  $\mathbf{w}_{c_t}$ .

To synthesize a sample *thin* from *his*, we query the substring *hi* and receive an array of DSDs:  $[\mathbf{w}_h, \mathbf{w}_{hi}]$ . As  $\mathbf{w}_t$  and  $\mathbf{w}_n$  are computed from  $\bar{\mathbf{w}}$ :

$$\mathbf{w}_t^{\text{rec}} = h_{\theta}([\mathbf{w}_t]) \quad (14a)$$

$$\mathbf{w}_{th}^{\text{rec}} = h_{\theta}([\mathbf{w}_t, \mathbf{w}_h]) \quad (14b)$$

$$\mathbf{w}_{thi}^{\text{rec}} = h_{\theta}([\mathbf{w}_t, \mathbf{w}_{hi}]) \quad (14c)$$

$$\mathbf{w}_{thin}^{\text{rec}} = h_{\theta}([\mathbf{w}_t, \mathbf{w}_{hi}, \mathbf{w}_n]) \quad (14d)$$

We use  $[\mathbf{w}_t, \mathbf{w}_{hi}]$  instead of  $[\mathbf{w}_t, \mathbf{w}_h, \mathbf{w}_{hi}]$  in Equations 14c and 14d because, as one might recall from generation Method  $\beta$  in the main paper (Sec. 3), the function approximator  $h_{\theta}$  is designed to *restore* temporal dependencies between writer-character-DSDs. As ‘h’ and ‘i’ are already temporally *dependent* within  $\mathbf{w}_{hi}$ , we need only connect characters ‘t’ and ‘h’ through LSTM  $h_{\theta}$ . The pseudocode for this sampling procedure is shown in Algorithm 1, with example generations in Figure 15.

**Input:**  $D$ : database of writer-character-DSD,  $s^*$ : target sentence to generate,  $\bar{\mathbf{w}}$ : mean global writer-DSD

```

1 Function PerformSamplingAlgorithm( $D, s^*, \bar{\mathbf{w}}$ ):
2   Initialize empty sets  $L, R$  and  $result$ 
3    $s^* \leftarrow \text{MarkAllCharactersAsUncovered}(s^*)$ 
4    $ss^* \leftarrow \text{ExtractSubStringsAndOrderByLength}(s^*)$ 
5   for each substring  $ss$  in  $ss^*$  do
6     if  $ss$  is in  $D$  and every characters in  $ss$  are not-covered then
7        $[\mathbf{w}_{c_1}, \dots, \mathbf{w}_{c_t}] = \text{QueryDatabaseWithKey}(ss)$ 
8       Add  $[\mathbf{w}_{c_1}, \dots, \mathbf{w}_{c_t}]$  to  $L$ 
9        $s^* \leftarrow \text{MarkCharactersInSubstringAsCovered}(s^*, ss)$ 
10  for each uncovered character  $c_t$  in  $s^*$  do
11     $\mathbf{w}_{c_t} \leftarrow \mathbf{C}_{c_t} \bar{\mathbf{w}}$ 
12    Add  $[\mathbf{w}_{c_t}]$  to  $L$ 
13   $L^* \leftarrow \text{OrderSetBySubstringAppearanceIn}(s^*)$ 
14  for each array  $A$  in  $L^*$  do
15    for each  $\mathbf{w}_{c_i}$  in  $A$  do
16       $\mathbf{w}_{c_i}^{\text{rec}} \leftarrow h_\theta([R_1, R_2, \dots, \mathbf{w}_{c_i}])$ 
17      Add  $\mathbf{w}_{c_i}^{\text{rec}}$  to the  $result$  list
18      if  $\mathbf{w}_{c_t}$  is the last element in  $A$  then
19        Add  $\mathbf{w}_{c_i}$  to the reference set  $R$ 
20  return  $result$ 

```

**Algorithm 1:** Pseudocode for our sampling algorithm to reconstruct writer-character-DSDs for the target sentence to synthesize.

## G Sequence Decoder $f_\theta^{\text{dec}}$

To synthesize a new sample from a list of writer-character-DSD  $\mathbf{w}_{c_t}$ , we train a sequence decoder function  $f_\theta^{\text{dec}}$ . The inputs to this decoder are: 1) initial point  $p_0 = (0, 0, 0)$ , and 2) the first writer-character-DSD  $\mathbf{w}_{c_1}$ . Continuing with the *thin* example, we predict the first point  $p_1$  from  $p_0$  and  $\mathbf{w}_t$ . At runtime, the predicted point  $p_1^*$  will be fed into the LSTM at the next timestep to predict  $p_2$ . When the decoder model outputs an  $eoc > 0.5$  (end-of-character probability), the model stops drawing the current character and start referencing the next writer-character-DSD so that it starts drawing the next character. This procedure is illustrated as the red lines in Figure 16. Similarly, to determine the touch/untouch status of the pen to the canvas, we use the  $eos$  (end-of-stroke probability) which is enclosed in point prediction  $p_t^*$ . If  $eos_t > 0.5$ , our model lifts up the pen; if  $eos_t \leq 0.5$ , our model continues the stroke.

Note that when we use the predicted  $p_t^*$  as an input to the LSTM at runtime, we binarize the  $eos$  value. This is because all  $eos$  values in training data are binarized. Further, we do not use the predicted points to predict the next point during training, because we have the true point sequence  $\mathbf{x}$ . In other words:

$$p_{t+1}^* = f_\theta^{\text{dec}}(p_0, p_1, \dots, p_t | \mathbf{w}_{c_t}) \quad (\text{training}) \quad (15a)$$

$$p_{t+1}^* = f_\theta^{\text{dec}}(p_0, p_1^*, \dots, p_t^* | \mathbf{w}_{c_t}) \quad (\text{runtime}) \quad (15b)$$

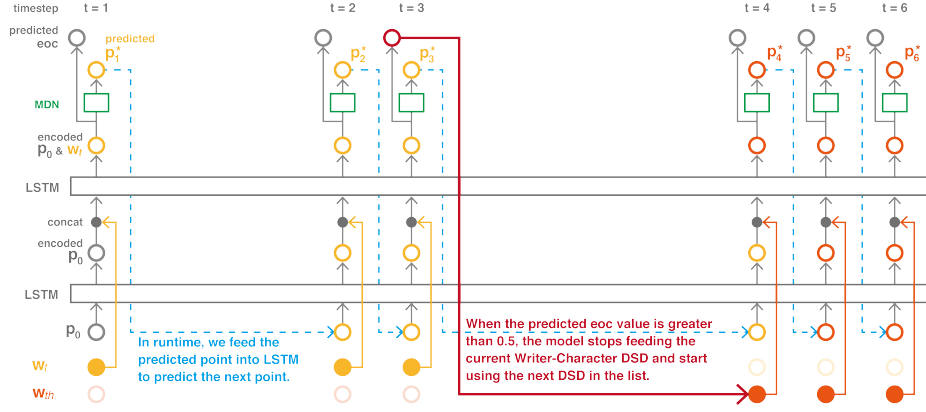


Fig. 16: Overview of our decoder architecture. During training, we feed true point sequences to the LSTM and do not use the predicted output  $p_t^*$  as the next input (the procedure shown as dotted blue lines).



Fig. 17: Variations in generated results from a single writer-character-DSD  $w_{c_t}$ , achieved by sampling points from predicted MDN distributions.

where  $*$  indicates *predicted* outputs by the decoder network.

Finally, the mixture density networks [10] (MDN) layer in our decoder makes it possible for our model to generate varying samples even from the same writer-character-DSD  $w_{c_t}$ . Examples are shown in Figure 17.

## H Character Encoder Function $g_\theta$

Next, we discuss in detail how the character matrix  $\mathbf{C}$  is computed. First, we convert each one-hot character vector  $\mathbf{c}_t$  in the sentence  $\mathbf{s}$  into a 256 dimensional vector  $\mathbf{c}_t^{\text{raw}}$  via a fully-connected layer  $g_\theta^{\text{FC1}}$ . Then, we feed this vector into LSTM  $g_\theta^{\text{LSTM}}$  and receive outputs  $\mathbf{c}_{c_t}^{\text{raw}}$  of the same size.  $g_\theta^{\text{LSTM}}$  is designed to encode temporal dependencies among characters. Then, we use a mapping function  $g_\theta^{\text{FC2}}$  to transform the  $256 \times 1$  vector into a 65,536 dimensional vector, and finally

reshape the output vector to a  $256 \times 256$  matrix  $\mathbf{C}_{c_t}$ . This process is as follows:

$$\mathbf{c}_t^{\text{raw}} = g_{\theta}^{\text{FC1}}(c_t) \quad (16a)$$

$$\mathbf{c}_{c_t}^{\text{raw}} = g_{\theta}^{\text{LSTM}}([\mathbf{c}_1^{\text{raw}}, \dots, \mathbf{c}_t^{\text{raw}}]) \quad (16b)$$

$$\mathbf{C}_{c_t} = \text{Reshape}(g_{\theta}^{\text{FC2}}(\mathbf{c}_{c_t}^{\text{raw}})) \quad (16c)$$

The parameters in  $g_{\theta}^{\text{FC2}}$  take up about one third of total number of parameters in our proposed model; this is expensive. However, using a fully-connected layer allows each value in the output  $\mathbf{C}_{c_t}$  to be computed from all values in the 256-dimensional vector  $\mathbf{c}_{c_t}^{\text{raw}}$ . If each value in  $\mathbf{c}_{c_t}^{\text{raw}}$  represents some different information about the character, then we intended to weight them 65,536 times via distinct mapping functions to create a matrix  $\mathbf{C}_{c_t}$ . We leave the study of other possible  $g_{\theta}$  architectures for future work.

## I Segmentation Network $k_{\theta}$

We introduce an unsupervised training technique to segment sequential handwriting samples into characters without any human intervention. For comparison, the existing state-of-the-art DeepWriting handwriting synthesis model [2] relies on commercial software for character segmentation.

Our data samples for training arrive as stroke sequences and character strings, with no explicit labeling on where one character ends and another begins within the stroke sequence. As such, we train a segmentation network  $k_{\theta}$  to segment sequential input data  $\mathbf{x}$  into characters, and to predict *end of character (eoc)* labels for each point in  $\mathbf{x}$ . Relying on these predicted *eoc* labels, we can extract  $\mathbf{w}_{c_t}$  from encoded  $\mathbf{x}^*$  and synthesize new samples with  $f_{\theta}^{\text{dec}}$ .

To prepare the input data for training, we extract 23 features per point  $p_t$  in  $\mathbf{x}$ , as is commonly used in previous work [19,29,31]. We feed these into a bidirectional LSTM to output a probability distribution over all character classes. From this output  $O$  of size  $(N, Q)$ , where  $N$  is the input sequence length and  $Q$  is the total number of characters, we compute a loss that is similar to a connectionist temporal classification (CTC) loss [18]. As seen in Figure 18, we make a slight modification in connections among nodes to adjust the change in two domains: character recognition and segmentation. In the recognition task, the blank character - was introduced to fill the gap between two character predictions (e.g.,  $a-b-b$ ), but because our goal is to label each point in the input sequence with a specific character in the corresponding sentence, we must avoid unnecessary use of the blank character and instead predict actual characters (e.g.,  $aaabbbb$ ). The only case where the blank character is needed in segmentation is when a character is repeated in a sentence. To highlight the switch from the first  $b$  case to the second  $b$  case, we use the - (e.g.,  $aaabb-b$ ). This slight modification in CTC connections enables us to train our segmentation network in an unsupervised manner, automatically label sequential handwriting data with characters and identify *eoc* indices. Examples of segmentation are shown in Figure 19.

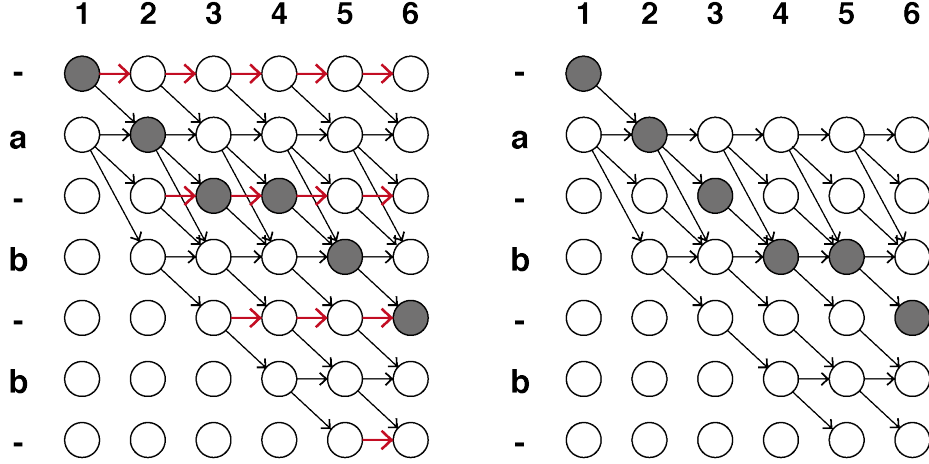


Fig. 18: Illustration of connections between temporal nodes. (Left) Original CTC connections. (Right) Our design of CTC connections. The connections between non-character nodes ‘-’ are prohibited (red arrows in the original). The shaded nodes shows an example route for prediction.

## J Detailed Training Procedure

### A Ablation Study

In the main paper, we discussed three different ways to obtain  $\mathbf{w}_{c_t}$ :  $f_{\theta}^{\text{enc}}$ , Method  $\alpha$ , and Method  $\beta$ . As we compute losses for each type, we conducted a simple ablation study. First, removing  $\mathcal{L}_{\alpha}$  from the total loss will take away the ability to generate handwriting samples from the mean writer-DSD  $\bar{\mathbf{w}}$  from the model by construction. Similarly, excluding  $\mathcal{L}_{\beta}$  will disallow our model to synthesize a new sample from saved writer-character-DSDs in the database  $D$  and only allow it to generate from the mean  $\bar{\mathbf{w}}$ . It is clear that we need both types of losses,  $\mathcal{L}_{\alpha}$  and  $\mathcal{L}_{\beta}$ , to have the current model capabilities.

However, eliminating  $\mathcal{L}_{f_{\theta}^{\text{enc}}}$ , the loss term for a method that uses the original writer-character-DSD extracted from  $f_{\theta}^{\text{enc}}$ , does not change model dynamics. Hence, we trained ablated models with modified loss functions that do not include: 1) any terms related to  $f_{\theta}^{\text{enc}}$  (i.e.,  $\mathcal{L}_{f_{\theta}^{\text{enc}}}$ ,  $\mathcal{L}_{\alpha}^{\mathbf{w}_{ct}}$  and  $\mathcal{L}_{\beta}^{\mathbf{w}_{ct}}$ ), and 2) just  $\mathcal{L}_{f_{\theta}^{\text{enc}}}$ . Figure 20 shows the training curve for word-level location losses  $\mathcal{L}^{\text{loc}}$ . Removing  $\mathcal{L}_{f_{\theta}^{\text{enc}}}$  had significant influence on Method  $\beta$  locational loss,  $\mathcal{L}_{\beta}^{\text{loc}}$  (standard:  $-3.213$  vs. ablated:  $-1.811$  after 250K training steps).

From this result, we assume that  $\mathcal{L}_{f_{\theta}^{\text{enc}}}$  works as a learning guideline for our model, and speeds up the training. We analyze that this is because having  $\mathcal{L}_{f_{\theta}^{\text{enc}}}$  in our loss function encourages accurate learning for our decoder function  $f_{\theta}^{\text{dec}}$ . In this setting, the function  $f_{\theta}$  is indeed an autoencoder, and the decoder is trained to restore  $\mathbf{x}$  from its encoded representation, writer-character-DSDs. This will

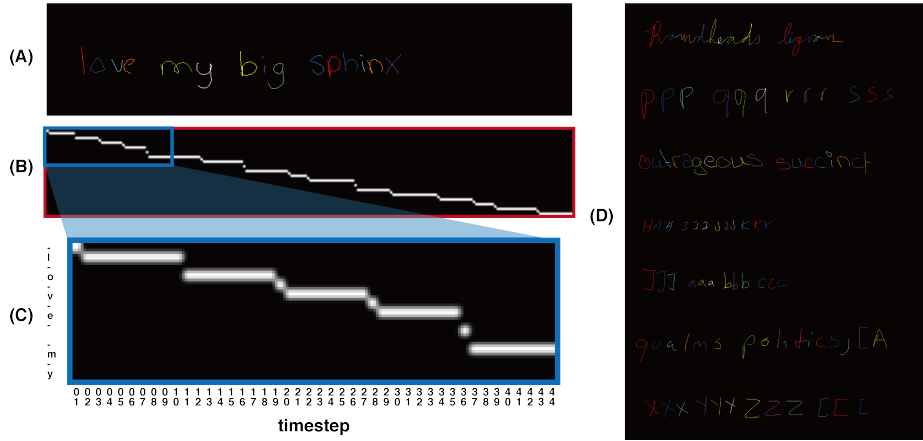


Fig. 19: Segmentation results. A) in handwriting image format. Different colors indicate different character segments. B) in CTC best route format. C) enlarged figure of the route path. D) More results.

increase the decoder performance, and as the decoder accuracy is maintained, the model can focus on learning the encoder problem, which is reconstruction of writer-character-DSDs by Method  $\alpha$  and  $\beta$ .

The reconstruction losses,  $\mathcal{L}_\alpha^{\text{w}_{ct}}$  and  $\mathcal{L}_\beta^{\text{w}_{ct}}$ , by contrast, did not affect the learning speed. We assume this can also be addressed by the same reason. Even if we constrained the reconstructed DSDs by Method  $\alpha$  and  $\beta$  to minimize their differences with the original DSDs from  $f_\theta^{\text{enc}}$ , those constraints will penalize the encoder more than they do for the decoder. To effectively train the decoder function, our model thus requires the loss term  $\mathcal{L}_{f_\theta^{\text{enc}}}$ .

## B Hyperparameters

To train our synthesis model, we use Adam [32] as our optimizer and set the learning rate to 0.001. We also clip the gradients in the range  $[-10.0, 10.0]$  to enhance learning stability. We use 5 sentence-level samples (relevant word-level and character-level samples are included as well) for each batch in training. We use multi-stacked (3-layers) LSTMs for our recurrent layers.

## K Dataset Specification and Collection Methodology

Our dataset considers the 86 characters: a space character ‘ ’, and the following 85 characters:

```
0123456789
abcdefghijklmnopqrstuvwxyz
ABCDEFGHIJKLMNOPQRSTUVWXYZ
!?"' *+-=: ; , . <> \ / [ ] ( ) # $ % &
```



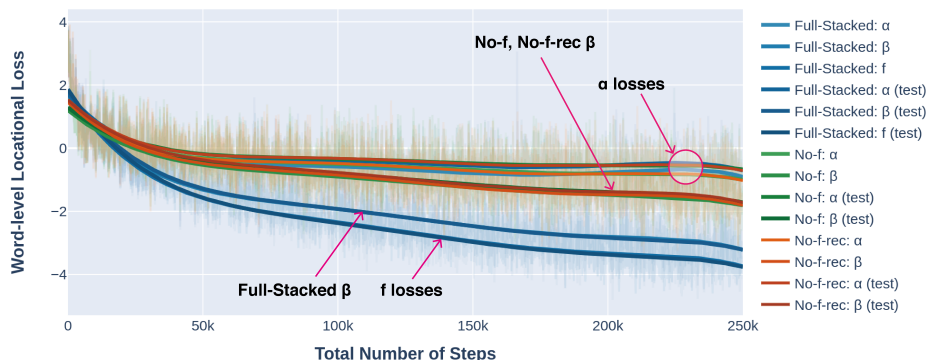


Fig. 20: Effects of  $\mathcal{L}_{f_{\theta}^{\text{enc}}}$  on training word-level location loss  $\mathcal{L}_{loc}^{\text{word}}$ . Transparent lines show the actual data points, and solid lines show smoothed training curves. *Full-Stacked model* is trained with the full loss, while *No-f model*'s loss function does not include  $\mathcal{L}_{f_{\theta}^{\text{enc}}}$ . Further, *No-f-rec model* does not have  $\mathcal{L}_{f_{\theta}^{\text{enc}}}$ ,  $\mathcal{L}_{\alpha}^{\text{w}_{ct}}$ ,  $\mathcal{L}_{\beta}^{\text{w}_{ct}}$  terms in its loss function. Test data is 20 held-out writers.

We collected handwriting samples from 170 writers using Amazon Mechanical Turk. An example screen of our data collection website is shown in Figure 21. Writing arbitrary words is laborious, and so we set a data-collection time limit of 60 minutes. Given this, it was necessary to select a subset of English words for our data collection.

## A Defining Target Words and Sentences

We analyzed the Gutenberg Dataset [35], which is a large corpus of 3,036 English books. These documents use 99 characters in total, including alphabetical, numerical, and special characters. In total, 5,831 character pairs appear in the dataset, while theoretically there are 9,702 possible character pairs ( $99 \times 99 - 99$ ). By counting the number of occurrences of each character pair, we constructed an ordered list of character pairs that is then used to score 2,158,445 distinctive words within the corpus.

The first word to be selected from the corpus was *therefore*, which includes the two most frequently used character pairs *th* and *he*. In fact, the character pairs within *therefore* appear so frequently that they altogether cover 13.5% of all character pair occurrences.

After adding *therefore* to the list of words for experiments, we then add additional words iteratively: we re-calculate scores for all other words with updated scores of character pairs (i.e., after adding *therefore*, the pairs *th* and *he* will not have high scores in future iterations). This process was repeated until the words in the list exceeded 99% coverage of all character pair occurrences.

Then, we constructed sentences from these high-scored words. Each sentence was less than 24 characters length to meet a space constraint due to our experiment

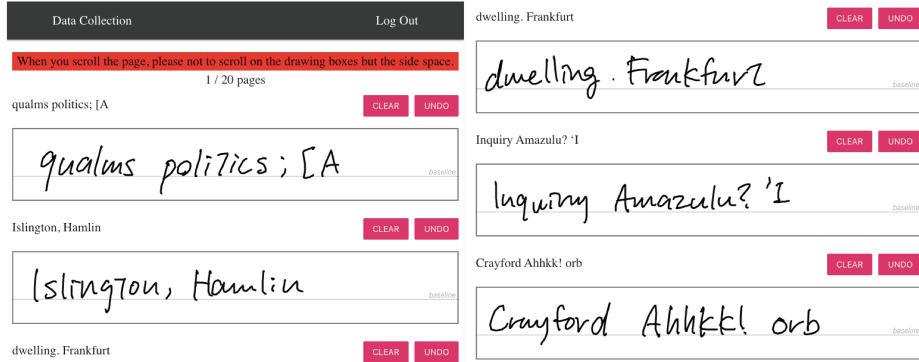


Fig. 21: Example screen of our data collection website. Each drawing box is 750 pixels  $\times$  120 pixels, and we provide a baseline at 80 pixels from the top.

setup. We asked tablet owners to write the prompted sentences using their stylus within the bounding box (750 pixels  $\times$  120 pixels), and 24 characters was the maximum number of characters that could reasonably fit into the region.

We also added several pangram sentences as well as repeated characters sentences (e.g., *aaa bbb ccc ddd*), and that led to our basic list of 192 sentences. These sentences use 86 unique characters, instead of 99 available characters, due to our decision to ignore rarely used special characters. They also use 1,182 distinct character pairs which cover 99.5% of all character pair occurrences (1,158,051,103/1,164,429,941). The remaining pairs could have been ignored, yet because that 0.5% was still large—6,378,838 occurrences by 4,649 character pairs—we decided to create a list of extra words with less frequently used character pairs, distribute them to 170 writers. Thus, each writer creates some rarer data that varies for each writer, in addition to their basic 192 sentences. As a result, we achieve 99.9% coverage with 3,894 character pairs.

## B Writer Behavior

Handwriting dataset collection is complex for various reasons, and in general creating a clean dataset without heuristic or manual cleaning is difficult. In our collection process, sometimes a writer would realize that s/he missed certain characters in the sentence after finishing the line, and so would go back to the location to add new strokes. These ‘late’ character additions are accidental rather than intentional. In contrast, conventional online handwriting recognition literature defines *delayed strokes*, where in cursive writing the horizontal bar of *t* and *f*, or the dot of *i* and *j*, are often added after a writer finished the current word. To distinguish between these two cases of late characters and delayed strokes, we disregard the temporal order of each stroke in a sample and reorder them from left to right *if* the leftmost point in a stroke is to the right of the rightmost point in another stroke. In this way, accidental omissions are removed.

Further, although we strongly advised participants to erase previous lines if they made mistakes, most participants either ignored this and left mistakes in, or scribbled over those regions to block them out. Writers also missed characters from the prompted sentences, and not a single participant (out of 170 writers) succeeded in near-perfect writing of 192 sentences. As our segmentation network (Sec. I) assumes that each drawing sample is labeled with the accurate character sequence, missed characters can directly affect the performance of segmentation. Hence, we manually corrected these instances throughout our dataset.