A Probabilistic Generative Model for Typographical Analysis of Early Modern Printing

Kartik Goyal[1] **Chris Dyer**[2] **Christopher Warren**[3] **Max G'Sell**[4] **Taylor Berg-Kirkpatrick**[5]
[1]Language Technologies Institute, Carnegie Mellon University   [2]Deepmind
[3]Department of English, Carnegie Mellon University
[4]Department of Statistics, Carnegie Mellon University
[5]Computer Science and Engineering, University of California, San Diego
{kartikgo,cnwarren,mgsell}@andrew.cmu.edu
cdyer@google.com   tberg@eng.ucsd.edu

Abstract

We propose a deep and interpretable probabilistic generative model to analyze glyph shapes in printed Early Modern documents. We focus on clustering extracted glyph images into underlying templates in the presence of multiple confounding sources of variance. Our approach introduces a neural editor model that first generates well-understood printing phenomena like spatial perturbations from template parameters via interpertable latent variables, and then modifies the result by generating a non-interpretable latent vector responsible for inking variations, jitter, noise from the archiving process, and other unforeseen phenomena associated with Early Modern printing. Critically, by introducing an inference network whose input is restricted to the visual residual between the observation and the interpretably-modified template, we are able to control and isolate what the vector-valued latent variable captures. We show that our approach outperforms rigid interpretable clustering baselines (Ocular) and overly-flexible deep generative models (VAE) alike on the task of *completely unsupervised* discovery of typefaces in mixed-font documents.

1 Introduction

Scholars interested in understanding details related to production and provenance of historical documents rely on methods of analysis ranging from the study of orthographic differences and stylometrics, to visual analysis of layout, font, and printed characters. Recently developed tools like Ocular (Berg-Kirkpatrick et al., 2013) for OCR of historical documents have helped automate and scale some textual analysis methods for tasks like compositor attribution (Ryskina et al., 2017) and digitization of historical documents (Garrette et al., 2015). However, researchers often find the need to go beyond
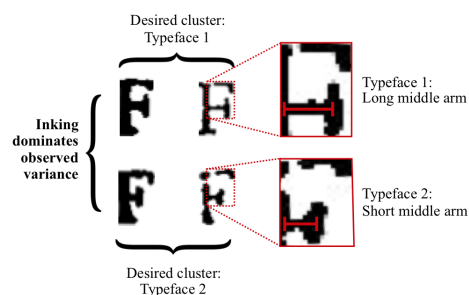

Figure 1: We desire a generative model that can be biased to cluster according to typeface characteristics (e.g. the length of the middle arm) rather than other more visually salient sources of variation like inking.

textual analysis for establishing provenance of historical documents. For example, Hinman (1963)'s study of typesetting in Shakespeare's First Folio relied on the discovery of pieces of damaged or distinctive type through manual inspection of every glyph in the document. More recently, Warren et al. (2020) examine pieces of distinctive types across several printers of the early modern period to posit the identity of clandestine printers of John Milton's *Areopagitica* (1644). In such work, researchers frequently aim to determine whether a book was produced by a single or multiple printers (Weiss (1992); Malcolm (2014); Takano (2016)). Hence, in order to aid these visual methods of analyses, we propose here a novel probabilistic generative model for analyzing extracted images of individual printed characters in historical documents. We draw from work on both deep generative modeling and interpretable models of the printing press to develop an approach that is both flexible and controllable – the later being a critical requirement for such analysis tools.

As depicted in Figure 1, we are interested in identifying clusters of subtly distinctive glyph shapes as these correspond to distinct metal stamps in the type-cases used by printers. However, other

sources of variation (inking, for example, as depicted in Figure 1) are likely to dominate conventional clustering methods. For example, powerful models like the variational autoencoder (VAE) (Kingma and Welling, 2014) capture the more visually salient variance in inking rather than typeface, while more rigid models (e.g. the emission model of Ocular (Berg-Kirkpatrick et al., 2013)), fail to fit the data. The goal of our approach is to account for these confounding sources of variance, while isolating the variables pertinent to clustering.

Hence, we propose a generative clustering model that introduces a neural editing process to add *expressivity*, but includes *interpretable* latent variables that model well-understood variance in the printing process: bi-axial translation, shear, and rotation of canonical type shapes. In order to make our model *controllable* and prevent deep latent variables from explaining all variance in the data, we introduce a restricted inference network. By only allowing the inference network to observe the visual residual of the observation after interpretable modifications have been applied, we bias the posterior approximation on the neural editor (and thus the model itself) to capture residual sources of variance in the editor – for example, inking levels, ink bleeds, and imaging noise. This approach is related to recently introduced neural editor models for text generation (Guu et al., 2018).

In experiments, we compare our model with rigid interpretable models (Ocular) and powerful generative models (VAE) at the task of unsupervised clustering subtly distinct typeface in scanned images early modern documents sourced from Early English Books Online (EEBO).

2 Model

Our model reasons about the printed appearances of a symbol (say majuscule F) in a document via a mixture model whose K components correspond to different metal stamps used by a printer for the document. During various stages of printing, random transformations result in varying printed manifestations of a metal cast on the paper. Figure 2 depicts our model. We denote an observed image of the extracted character by X. We denote choice of typeface by latent variable c (the mixture component) with prior π. We represent the shape of the k-th stamp by template T_k, a square matrix of parameters. We denote the interpretable latent variables corresponding to spatial adjustment of
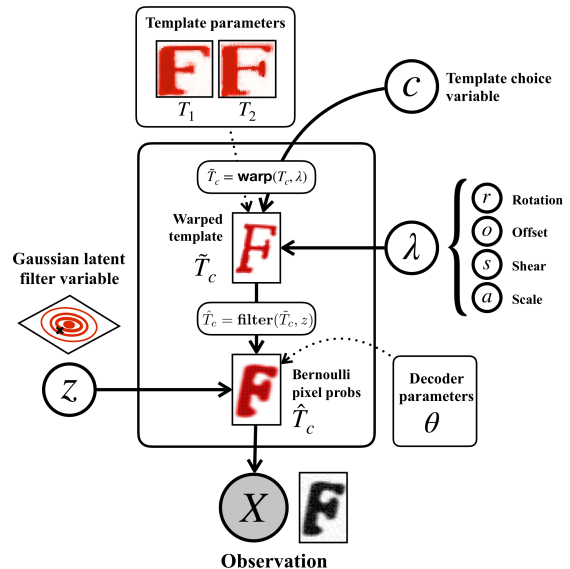

Figure 2: Proposed generative model for clustering images of a symbol by typeface. Each mixture component c corresponds to a *learnable* template T_k. The λ variables warp (spatially adjust) the original template T to $\tilde{T}$. This warped template is then further transformed via the z variables to $\hat{T}$ via an expressive neural filter function parametrized by θ.

the metal stamp by λ, and the editor latent variable responsible for residual sources of variation by z. As illustrated in Fig. 2, after a cluster component $c = k$ is selected, the corresponding template T_k undergoes a transformation to yield $\hat{T}_k$. This transformation occurs in two stages: first, the interpretable spatial adjustment variables (λ) produce an adjusted template (§2.1), $\tilde{T}_k = \text{warp}(T_k, \lambda)$, and then the neural latent variable transforms the adjusted template (§2.2), $\hat{T}_k = \text{filter}(\tilde{T}_k, z)$. The marginal probability under our model is

$$p(X) = \sum_k \pi_k \int p(X|\lambda, z; T_k)p(\lambda)p(z)dzd\lambda,$$

where $p(X|\lambda, z; T_k)$ refers to the distribution over the binary pixels of X where each pixel has a bernoulli distribution parametrized by the value of the corresponding pixel-entry in $\hat{T}_k$.

2.1 Interpretable spatial adjustment

Early typesetting was noisy, and the metal pieces were often arranged with slight variations which resulted in the printed characters being positioned with small amounts of offset, rotation and shear. These real-valued spatial adjustment variables are denoted by $\lambda = (r, o, s, a)$, where r represents the rotation variable, $o = (o_h, o_v)$ represents offsets along the horizontal and vertical axes, $s = (s_h, s_v)$

denotes shear along the two axes. A scale factor, $\tilde{a} = 1.0 + a$, accounts for minor scale variations arising due to the archiving and extraction processes. All variables in λ are generated from a Gaussian prior with zero mean and fixed variance as the transformations due to these variables tend to be subtle.

In order to incorporate these deterministic transformations in a differentiable manner, we map λ to a template sized attention map H_{ij} for each output pixel position (i, j) in $\tilde{T}$ as depicted in Figure 3. The attention map for each output pixel is formed in order to attend to the corresponding shifted (or scaled or sheared) portion of the input template and is shaped according to a Gaussian distribution with mean determined by an affine transform. This approach allows for strong inductive bias which contrasts with related work on *spatial-VAE* (Bepler et al., 2019) that learns arbitrary transformations.
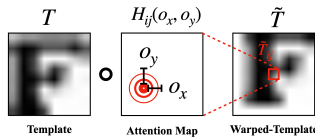

Figure 3: Translation operation: The mode of the attention map is shifted by the offset values for every output pixel in $\tilde{T}$. Similar operations account for shear, rotation, and scale.

2.2 Residual sources of variations

Apart from spatial perturbations, other major sources of deviation in early printing include random inking perturbations caused by inconsistent application of the stamps, unpredictable ink bleeds, and noise associated with digital archiving of the documents. Unlike in the case of spatial perturbations which could be handled by deterministic affine transformation operators, it is not possible to analytically define a transformation operator due to these variables. Hence we propose to introduce a non-interpretable real-valued latent vector z, with a Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$, that transforms $\tilde{T}$ into a final template $\hat{T}$ via neurally-parametrized function filter$(\tilde{T}, z; \theta)$ with neural network parameters θ. This function is a convolution over $\tilde{T}$ whose kernel is parametrized by z, followed by non-linear operations. Intuitively, parametrizing the filter by z results in the latent variable accounting for variations like inking appropriately because convolution filters capture local variations in appearance. Srivatsan et al. (2019) also observed the effectiveness of using z to define a deconvolutional kernel for
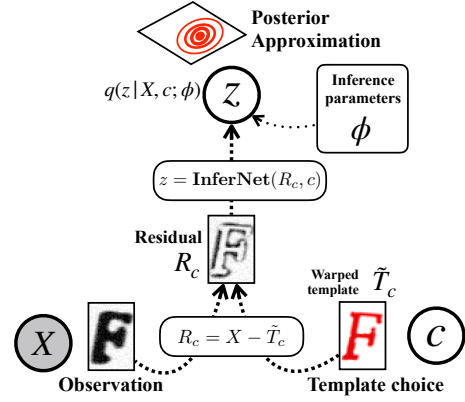

Figure 4: Inference network for z conditions on the mixture component and only the residual image left after subtracting the λ-transformed template from the image. This encourages z to model variance due to sources other than spatial adjustments.

font generation.

2.3 Learning and Inference

Our aim is to maximize the log likelihood of the observed data ($\{X_d \mid d \in \mathbb{N}, d < n\}$) of n images wrt. model parameters:

$$\mathrm{LL}(T_{1,\dots,k}, \theta) = \max_{T,\theta} \sum_d \log \Big[\sum_k \pi_k$$

$$\int p(X_d|\lambda_d, z_d; T_k, \theta) p(\lambda_d) p(z_d) dz_d d\lambda_d \Big]$$

During training, we maximize the likelihood wrt. λ instead of marginalizing, which is an approximation inspired by iterated conditional modes (Besag, 1986):

$$\max_{T,\theta} \sum_d \log \sum_k \max_{\gamma_{k,d}} \pi_k \int p(X_d|\lambda_d = \gamma_{k,d}, z_d;$$

$$T_k, \theta) p(\lambda_d = \gamma_{k,d}) p(z_d) dz_d$$

However, marginalizing over z remains intractable. Therefore we perform amortized variational inference to define and maximize a lower bound on the above objective (Kingma and Welling, 2014). We use a convolutional inference neural network parametrized by ϕ (Fig. 4), that takes as input, the mixture component k, the residual image $R_k = X - \tilde{T}_k$, and produces mean and variance parameters for an isotropic gaussian proposal distribution $q(z \mid R_k, k; \phi)$. This results in the final training objective:

$$\max_{T,\theta,\phi} \sum_d \log \sum_k \mathrm{E}_{q(z_d|R_{d,k},k;\phi)} \Big[\max_{\gamma_{k,d}} \big(\pi_k$$

$$p(X_d|\lambda = \gamma_{k,d}, z_d; T_k, \theta) p(\lambda = \gamma_{k,d}) \big) \Big]$$

$$- \mathrm{KL}\big(q(z_d|R_{d,k}, k; \phi) || p(z) \big)$$

We use stochastic gradient ascent to maximize this objective with respect to T, γ, θ and ϕ.

3   Experiments

We train our models on printed occurrences of 10 different uppercase character classes that scholars have found useful for bibliographic analysis (Warren et al., 2020) because of their distinctiveness. As a preprocessing step, we ran Ocular (Berg-Kirkpatrick et al., 2013) on the grayscale scanned images of historical books in EEBO dataset and extracted the estimated image segments for the letters of interest.

3.1   Quantitative analysis

We show that our model is superior to strong baselines at clustering subtly distinct typefaces (using realistic synthetic data), as well as in terms of fitting the real data from historical books.

3.1.1   Baselines for comparison

Ocular: Based on the emission model of Ocular that uses discrete latent variables for the vertical/horizontal offset and inking variables, and hence has limited expressivity.

λ-only: This model only has the interpretable continuous latent variables pertaining to spatial adjustment.

VAE-only: This model is expressive but doesn't have any interpretable latent variables for explicit control. It is an extension of Kingma et al. (2014)'s model for semi-supervised learning with a continuous latent variable vector in which we obtain tighter bounds by marginalizing over the cluster identities explicitly. For fair comparison, the encoder and decoder convolutional architectures are the same as the ones in our full model. The corresponding training objective for this baseline is:

$$\max_{T,\theta,\phi} \sum_d \log \sum_k \mathrm{E}_{q(z_d|X_d,k;\phi)}\big[\pi_k p(X_d|z_d; T_k, \theta)\big]$$
$$-\mathrm{KL}\big(q(z_d|X_d, k; \phi)||p(z)\big)$$

No-residual: The only difference from the full model is that the encoder for the inference network conditions the variational distribution $q(z)$ on the entire input image X instead of just the residual image $X - \tilde{T}$.

3.1.2   Font discovery in Synthetic Data

Early modern books were frequently composed from two or more type cases, resulting in documents with mixed fonts. We aim to learn the dif-

	V-measure	Mutual Info	F&M	NLL
Ocular	0.42	0.45	0.61	379.21
λ-only	0.49	0.51	0.70	322.04
VAE-only	0.22	0.29	0.38	263.45
No-residual	0.54	0.58	0.73	264.27
Our Model	**0.73**	**0.74**	**0.85**	**257.92**

Table 1: (a) Clustering results on synthetic data (**V-measure**, **Mutual Info**, **F&M**). (b) Test negative log likelihood (NLL) on real data from historical documents, or negative ELBO bound for intractable models (**NLL**).

ferent shapes of metal stamps that were used as templates for each cluster component in our model.

Data: In order to quantitatively evaluate our model's performance, we experiment with synthetically generated realistic dataset for which we know the ground truth cluster identities in the following manner: For each character of interest, we pick three distinct images from scanned segmented EEBO images, corresponding to three different metal casts. Then we randomly add spatial perurbations related to scale, offset, rotation and shear. To incorporate varying inking levels and other distortions, we randomly either perform *erosion*, *dilation*, or a combination of these warpings using OpenCV (Bradski, 2000) with randomly selected kernel sizes. Finally, we add a small Gaussian noise to the pixel intensities and generate 300 perturbed examples per character class.

Results: We report macro-averaged results across all the character classes on three different clustering measures, V-measure (Rosenberg and Hirschberg, 2007), Mutual Information and Fowlkes and Mallows Index (Fowlkes and Mallows, 1983). In Table 1, we see that our model significantly outperforms all other baselines on every metric. *Ocular* and *λ-only* models fail because they lack expressiveness to explain the variations due to random jitters, erosions and dilations. The *VAE-only* model, while very expressive, performs poorly because it lacks the inductive bias needed for successful clustering. The *No-residual* model performs decently but our model's superior performance emphasizes the importance of designing a restrictive inference network such that z only focuses on extraneous sources of variation.

3.1.3   Fitting Real Data from Historical Books

For the analysis of real books, we selected three books from the EEBO dataset printed by different printers. We modeled each character class for each book separately and report the macro-aggregated

upper bounds on the negative log likelihood (NLL) in Table 1. We observe that adding a small amount of expressiveness makes our λ-*only* model better than *Ocular*. The upper bounds of other inference network based models are much better than the tight[1] bounds of both the interpretable models. Our model has the lowest upper bound of all the models while retaining interpretability and control.

3.2 Qualitative analysis

We provide visual evidence of desirable behavior of our model on collections of character extractions from historical books with mixed fonts. Specifically, we discus the performance of our model on the mysterious edition of Thomas Hobbes' *Leviathan* known as "the 25 Ornaments" edition. (Hobbes, 1651 [really 1700?]). The 25 Ornaments *Leviathan* is an interesting test case for several reasons. While its title page indicates a publisher and year of publication, both are fabricated (Malcolm, 2014). The identities of its printer(s) remain speculative, and the actual year of publication is uncertain. Further, the 25 Ornaments exhibits two distinct fonts.

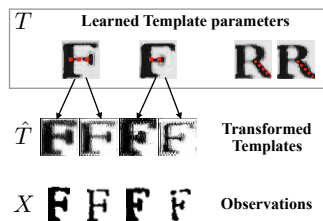### 3.2.1 Quality of learned templates

Figure 5: The learned templates for **F** and **R** and the transformed templates $\hat{T}$ for four examples of **F** are shown. Our model is able to learn desirable templates based on underlying glyph structure.

Our model is successful in discovering distinctly shaped typefaces in the 25 Ornaments *Leviathan*. We focus on the case study of majuscule letters **F** and **R**, each of which have two different typefaces mixed in throughout. The two typefaces for **F** differ in the *length of the middle arm* (Fig. 1), and the two typefaces for **R** have *differently shaped legs*. In Fig. 5, we show that our model successfully learns the two desired templates T_1 and T_2 for both the characters which indicates that the clusters in our

model mainly focus on subtle differences in underlying glyph shapes. We also illustrate how the latent variables transform the model templates T to $\hat{T}$ for four example **F** images. The model learns complex functions to transform the templates which go beyond simple affine and morphological transformations in order to account for inking differences, random jitter, contrast variations etc.

3.2.2 Interpretable variables (λ) and Control

Figure 6: Result of alignment on *Leviathan* extractions using the interpretable λ variables along with their pixelwise average images. Aligned average image is much sharper than the unaligned average image.

Finally, we visualize the ability of our model to separate responsibility of modelling variation among the interpretable and non-interpretable variables appropriately. We use the inferred values of the interpretable (λ) variable for each image in the dataset to adjust the corresponding image. Since the templates represent the canonical shape of the letters, the λ variables which shift the templates to explain the images can be *reverse applied* to the input images themselves in order to *align* them by accounting for offset, rotation, shear and minor size variations. In Fig. 6, we see that the input images (top row) are uneven and vary by size and orientation. By reverse applying the inferred λ values, we are able to project the images to a fixed size such that they are aligned and any remaining variations in the data are caused by other sources of variation. Moreover, this *alignment* method would be crucial for automating certain aspects of bibliographic studies that focus on comparing specific imprints.

4 Conclusion

Beyond applications to typeface clustering, the general approach we take might apply more broadly to other clustering problems, and the model we developed might be incorporated into OCR models for historical text.

5 Acknowledgements

---

[1]For **Ocular** and λ-**only** models, we report the upper bound obtained via maximization over the interpretable latent variables. Intuitively, these latent variables are likely to have unimodal posterior distributions with low variance, hence this approximation is likely tight.

References

Tristan Bepler, Ellen Zhong, Kotaro Kelley, Edward Brignole, and Bonnie Berger. 2019. Explicitly disentangling image content from translation and rotation with spatial-vae. In *Advances in Neural Information Processing Systems*, pages 15409–15419.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 207–217, Sofia, Bulgaria. Association for Computational Linguistics.

Julian Besag. 1986. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):259–279.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Edward B Fowlkes and Colin L Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.

Dan Garrette, Hannah Alpert-Abrams, Taylor Berg-Kirkpatrick, and Dan Klein. 2015. Unsupervised code-switching for multilingual historical document transcription. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1036–1041, Denver, Colorado. Association for Computational Linguistics.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Charlton Hinman. 1963. *The printing and proofreading of the first folio of Shakespeare*, volume 1. Oxford: Clarendon Press.

Thomas Hobbes. 1651 [really 1700?]. *Leviathan, or, the matter, form, and power of a common-wealth ecclesiastical and civil. By Thomas Hobbes of Malmesbury*. Number R13935 in ESTC. [false imprint] printed for Andrew Crooke, at the Green Dragon in St. Pauls Church-yard, London.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589.

Noel Malcolm. 2014. Editorial Introduction. In *Leviathan*, volume 1. Clarendon Press, Oxford.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

Maria Ryskina, Hannah Alpert-Abrams, Dan Garrette, and Taylor Berg-Kirkpatrick. 2017. Automatic compositor attribution in the first folio of shakespeare. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 411–416, Vancouver, Canada. Association for Computational Linguistics.

Nikita Srivatsan, Jonathan Barron, Dan Klein, and Taylor Berg-Kirkpatrick. 2019. A deep factorization of style and structure in fonts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2195–2205, Hong Kong, China. Association for Computational Linguistics.

Akira Takano. 2016. Thomas Warren: A Printer of Leviathan (head edition). *Annals of Nagoya University Library Studies*, 13:1–17.

Christopher N. Warren, Pierce Williams, Shruti Rijhwani, and Max G'Sell. 2020. Damaged type and Areopagitica's clandestine printers. *Milton Studies*, 62.1.

Adrian Weiss. 1992. Shared Printing, Printer's Copy, and the Text(s) of Gascoigne's "A Hundreth Sundrie Flowres". *Studies in Bibliography*, 45:71–104.

A Character wise quantitative analysis

The quantitative experiments were performed on the following character classes: A, B, E, F, G, H, M, N, R, W.

	V-measure	Mutual Info	F&M	NLL
λ-only	0.77	0.82	0.89	264.90
VAE-only	0.33	0.38	0.5	230.45
No-residual	0.79	0.85	0.90	231.45
Our Model	**0.78**	**0.86**	**0.89**	**226.25**

Table 2: Results for character A

	V-measure	Mutual Info	F&M	NLL
λ-only	0.37	0.39	0.59	261.1
VAE-only	0.15	0.2	0.32	229.1
No-residual	0.37	0.39	0.58	228.1
Our Model	**0.68**	**0.73**	**0.81**	**226.25**

Table 3: Results for character B

	V-measure	Mutual Info	F&M	NLL
λ-only	0.33	0.36	0.55	282.4
VAE-only	0.17	0.19	0.30	253.2
No-residual	0.33	0.35	0.56	251.45
Our Model	**0.65**	**0.70**	**0.76**	**234.05**

Table 4: Results for character E

	V-measure	Mutual Info	F&M	NLL
λ-only	0.65	0.71	0.79	418.01
VAE-only	0.31	0.45	0.42	364.2
No-residual	0.72	0.78	0.82	369.5
Our Model	**0.72**	**0.79**	**0.84**	**364.21**

Table 11: Results for character W

	V-measure	Mutual Info	F&M	NLL
λ-only	0.09	0.10	0.55	258.40
VAE-only	0.03	0.05	0.31	218.2
No-residual	0.12	0.09	0.59	208.1
Our Model	**0.81**	**0.56**	**0.94**	**204.48**

Table 5: Results for character F

	V-measure	Mutual Info	F&M	NLL
λ-only	0.60	0.62	0.73	268.40
VAE-only	0.28	0.38	0.40	250.8
No-residual	**0.64**	**0.66**	**0.77**	244.5
Our Model	0.60	0.62	0.73	240.84

Table 6: Results for character G

	V-measure	Mutual Info	F&M	NLL
λ-only	0.72	0.71	0.79	313.75
VAE-only	0.32	0.32	0.40	254.2
No-residual	0.90	0.97	0.94	258.8
Our Model	**0.92**	**1.01**	**0.96**	**249.81**

Table 7: Results for character H

	V-measure	Mutual Info	F&M	NLL
λ-only	0.62	0.64	0.78	392.06
VAE-only	0.29	0.38	0.40	323.5
No-residual	0.70	0.83	0.74	329.25
Our Model	**0.75**	**0.84**	**0.87**	**323.04**

Table 8: Results for character M

	V-measure	Mutual Info	F&M	NLL
λ-only	0.65	0.70	0.73	331.6
VAE-only	0.30	0.45	0.40	265.2
No-residual	**0.74**	**0.81**	**0.82**	270.11
Our Model	0.69	0.75	0.75	**264.23**

Table 9: Results for character N

	V-measure	Mutual Info	F&M	NLL
λ-only	0.07	0.08	0.55	330.6
VAE-only	0.03	0.04	0.34	247.1
No-residual	0.06	0.07	0.53	251.32
Our Model	**0.46**	**0.32**	**0.78**	**246.02**

Table 10: Results for character R