Morpho-MNIST: Quantitative Assessment and Diagnostics for Representation Learning

Daniel C. Castro   DC315@IMPERIAL.AC.UK

Jeremy Tan   JHT16@IMPERIAL.AC.UK

Bernhard Kainz   B.KAINZ@IMPERIAL.AC.UK
Biomedical Image Analysis Group
Imperial College London
London SW7 2AZ, United Kingdom

Ender Konukoglu   ENDER.KONUKOGLU@VISION.EE.ETHZ.CH
Computer Vision Laboratory
ETH Zürich
8092 Zürich, Switzerland

Ben Glocker   B.GLOCKER@IMPERIAL.AC.UK
Biomedical Image Analysis Group
Imperial College London
London SW7 2AZ, United Kingdom

Editor: Isabelle Guyon

Abstract

Revealing latent structure in data is an active field of research, having introduced exciting technologies such as variational autoencoders and adversarial networks, and is essential to push machine learning towards unsupervised knowledge discovery. However, a major challenge is the lack of suitable benchmarks for an objective and quantitative evaluation of learned representations. To address this issue we introduce Morpho-MNIST, a framework that aims to answer: "to what extent has my model learned to represent specific factors of variation in the data?" We extend the popular MNIST dataset by adding a morphometric analysis enabling quantitative comparison of trained models, identification of the roles of latent variables, and characterisation of sample diversity. We further propose a set of quantifiable perturbations to assess the performance of unsupervised and supervised methods on challenging tasks such as outlier detection and domain adaptation. Data and code are available at `https://github.com/dccastro/Morpho-MNIST`.

Keywords: representation learning, generative models, empirical evaluation, disentanglement, morphometrics

1. Introduction

A key factor for progress in machine learning has been the availability of well curated, easy-to-use, standardised and sufficiently large annotated datasets for benchmarking different algorithms and models. This has led to major advances in speech recognition, computer vision, and natural language processing. A commonality between these tasks is their natural

formulation as supervised learning tasks, wherein performance can be measured in terms of accuracy on a test set.

The general problem of representation learning (i.e. to reveal latent structure in data) is more difficult to assess due the lack of suitable benchmarks. Although the field is very active, with many recently proposed techniques such as probabilistic autoencoders and adversarial learning, it is less clear where the field stands in terms of progress or which approaches are more expressive for specific tasks. The lack of reproducible ways to quantify performance has led to subjective means of evaluation: visualisation techniques have been used to show low-dimensional projections of the latent space and visual inspection of generated or reconstructed samples are popular to provide subjective measures of descriptiveness. On the other hand, the quality of sampled images generally tells us little about how well the learned representations capture known factors of variation in the training distribution. In order to advance progress, the availability of tools for objective assessment of representation learning methods seems essential yet lacking.

This paper introduces Morpho-MNIST, a collection of shape metrics and perturbations, in a step towards quantitative assessment of representation learning. We build upon one of the most popular machine learning benchmarks, MNIST, which despite its shortcomings remains widely used. While MNIST was originally constructed to facilitate research in image classification, in the form of recognising handwritten digits (LeCun et al., 1998), it has found its use in representation learning, for example, to demonstrate that the learned latent space yields clusters consistent with digit labels. Methods aiming to disentangle the latent space claim success if individual latent variables capture specific style variations (e.g. stroke thickness, sidewards leaning digits and other visual characteristics).

It is important to emphasise that MNIST is a prime representative for datasets of rasterised two-dimensional shapes. In this type of data, while the underlying shapes may have relatively simple and well understood factors of variation, they manifest in the raster images as high-order correlations between large numbers of pixels. The challenge for a learning agent is to abstract away the low-level pixel intensity variations and recover a meaningful shape representation. MNIST is one of the interesting cases wherein it is possible to explicitly measure morphological attributes from the images themselves, enabling us to directly evaluate to what extent a trained model has learned to represent them.

More generally, MNIST remains popular in practice due to several factors: it allows reproducible comparisons with previous results reported in the literature; the dataset is sufficiently large for its complexity and consists of small, two-dimensional greyscale images defining a tractable ten-class classification problem; computation and memory requirements are low; most popular deep learning frameworks and libraries offer tutorials using MNIST, which makes it straightforward for new researchers to enter the field and to experiment with new ideas and explore latest developments. Furthermore, there is now renewed interest in MNIST due to the recent replication of the original generation pipeline and the rediscovery of additional 50,000 test images, beckoning researchers to re-evaluate over 20 years of research on MNIST (Yadav and Bottou, 2019). We leverage these qualities and extend MNIST in multiple ways, as summarised in the following.
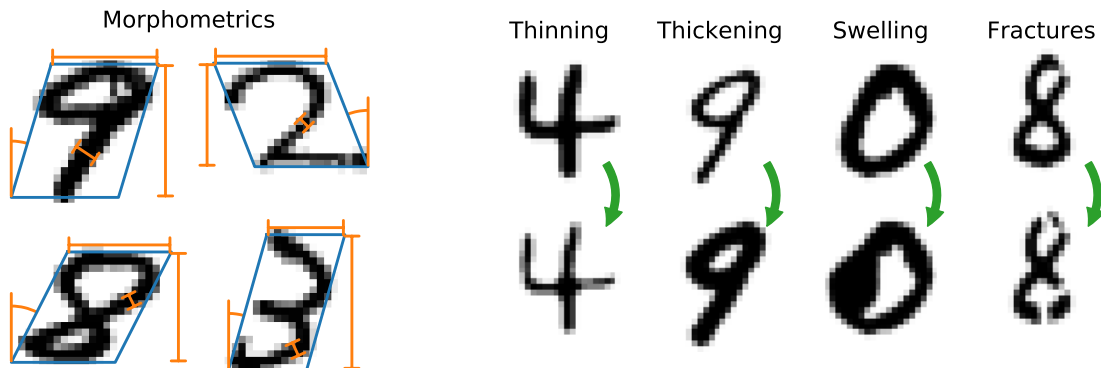
Figure 1: *Left:* MNIST morphometrics—stroke thickness and length (not shown), width, height and slant of digits. *Right:* MNIST perturbations (many more examples of each type in Appendix B).

1.1. Contributions

Our aim is to provide a useful resource to the machine learning community, opening a variety of new avenues for experimentation and analysis. As we preserve the general properties of MNIST—such as image size, file format, numbers of training and test images, and the original ten-class classification problem—we hope this new quantitative framework for assessing representation learning will experience widespread adoption and may inspire further extensions facilitated by a publicly available Morpho-MNIST code base.

1.1.1. MORPHOMETRICS

We propose to describe real and generated digit images in terms of measurable shape attributes. These include stroke thickness and length, and the width, height, and slant of digits (see Fig. 1, left). Whereas some of these MNIST properties have been analysed only visually in previous work, we demonstrate that quantifying each of them allows us to objectively characterise the roles of inferred representations.

More generally, the exclusive reliance on synthetic annotated datasets (Section 1.2.4) has been a crucial limitation for research into evaluation of learned representations (Eastwood and Williams, 2018), as this restricts us to the analysis of *inferential* behaviour (data $\rightarrow$ representation), even when studying generative models. The tools we introduce here can be used to measure model samples, effectively enabling direct assessment of *generative* performance (representation $\rightarrow$ data).

These measurements can be directly employed to re-evaluate existing models and may be added retrospectively to previous experiments involving the original MNIST dataset. Adoption of our morphometric analysis may provide new insights into the effectiveness of representation learning methods in terms of revealing meaningful latent structures. Furthermore, for other datasets it suffices to design the relevant scalar metrics and include them in the very same evaluation framework.

1.1.2. Perturbations

We introduce a set of parametrisable global and local perturbations, inspired by natural and pathological variability in medical images. Global changes involve overall thinning and thickening of digits, while local changes include both swelling and fractures (see examples on the right in Fig. 1 and many more in Appendix B). The perturbation framework we present in Section 3—illustrating various ways to exploit the outputs of the proposed processing pipeline (e.g. skeleton, distance map, and even morphometrics themselves; see Section 2.1)— is sufficiently general that additional transformations can easily be designed as necessary. Injecting these perturbations into the dataset adds a new type of complexity to the data manifold and opens up a variety of interesting applications.

The proposed perturbations are designed to enable a wide range of new studies and applications for both supervised and unsupervised tasks. Detection of 'abnormalities' (i.e. local perturbations) is an evident application, although more challenging tasks can also be defined, such as classification from noisy/corrupted data, domain adaptation, localisation of perturbations, characterising semantics of learned latent representations, and more. We explore a few supplementary examples of supervised tasks in Section 4.3.

1.2. Related Work: Datasets

In this section, we provide an overview of some datasets that are related to MNIST, by either sharing its original source content, containing transformations of the original MNIST images or being distributed in the same format for easy replacement. We also mention a few prevalent datasets of synthetic images with generative factor annotations, analogous to the morphometrics proposed in this paper.

1.2.1. NIST Datasets

The MNIST (modified NIST) dataset (LeCun et al., 1998) was constructed from handwritten digits in NIST Special Databases 1 and 3, now released as Special Database 19 (Grother and Hanaoka, 2016). Cohen et al. (2017) generated a much larger dataset based on the same NIST database, containing additional upper- and lower-case letters, called EMNIST (extended MNIST). LeCun et al.'s original processing pipeline was recently replicated, enabling the rediscovery of 50,000 test images that had been excluded from MNIST—a dataset now released as QMNIST (Yadav and Bottou, 2019).

1.2.2. MNIST Perturbations

The seminal paper by LeCun et al. (1998) employed data augmentation using planar affine transformations including translation, scaling, squeezing, and shearing. Loosli et al. (2007) employed random elastic deformations to construct the Infinite MNIST dataset. Other MNIST variations include rotations and insertion of random and structured background (Larochelle et al., 2007), and Tieleman (2013) applied spatial affine transformations and provided ground-truth transformation parameters.

1.2.3. MNIST Format

Due to the ubiquity of the MNIST dataset in machine learning research and the resulting multitude of compatible model architectures available, it is appealing to release new datasets in the same format (28×28, 8-bit grayscale images). One such effort is Fashion-MNIST (Xiao et al., 2017), containing images of clothing articles from ten distinct classes, adapted from an online shopping catalogue. Another example is notMNIST (Bulatov, 2011), a dataset of character glyphs for letters 'A'–'J' (also ten classes), in a challengingly diverse collection of typefaces.

1.2.4. Annotated Datasets

Synthetic computer vision datasets that are popular for evaluating disentanglement of learned latent factors of variation include those from Paysan et al. (2009) and Aubry et al. (2014). They contain 2D renderings of 3D faces and chairs, respectively, with ground-truth pose parameters (azimuth, elevation) and lighting conditions (faces only). A further initiative in that direction is the dSprites dataset (Matthey et al., 2017), which consists of binary images containing three types of shapes with varying location, orientation and size. The availability of the ground-truth values of such attributes has motivated the accelerated adoption of these datasets in the evaluation of representation learning algorithms.

1.3. Related Work: Quantitative Evaluation

Evaluation of representation learning is a notoriously challenging task and remains an open research problem. Numerous solutions have been proposed, with many of the earlier ones focusing on the test log-likelihood under the model (Kingma and Welling, 2014) or, for likelihood-free models, under a kernel density estimate (KDE) of generated samples (Goodfellow et al., 2014; Makhzani et al., 2016)—shown not to be reliable proxies for the true model likelihood (Theis et al., 2016).

Another perspective for evaluation of generative models of images is the visual fidelity of its samples to the training data, which would normally require manual inspection. To address this issue, a successful family of metrics have been proposed, based on visual features extracted by the Inception network (Szegedy et al., 2016). The original Inception score (Salimans et al., 2016) relies on the 'crispness' of class predictions, whereas the Fréchet Inception distance (FID, Heusel et al., 2017) and the kernel Inception distance (KID, Bińkowski et al., 2018) statistically compare high-level representations instead of the final network outputs.

Although the approaches above can reveal vague signs of mode collapse, it may be useful to diagnose this phenomenon on its own. With this objective, Arora et al. (2018) proposed to estimate the support of the learned distribution (assumed discrete) using the birthday paradox test, by counting pairs of visual duplicates among model samples. Unfortunately, the adoption of this technique is hindered by its reliance on manual visual inspection to flag identical images.

There have been several attempts at quantifying representation disentanglement performance. For example, Higgins et al. (2017) proposed to use the accuracy in predicting which factor of variation was held fixed in a simulated dataset. There exist further information-theoretic approaches, involving the KL divergence contribution from each latent dimension (Dupont, 2018) or their mutual information with each known generative factor (Chen et al.,
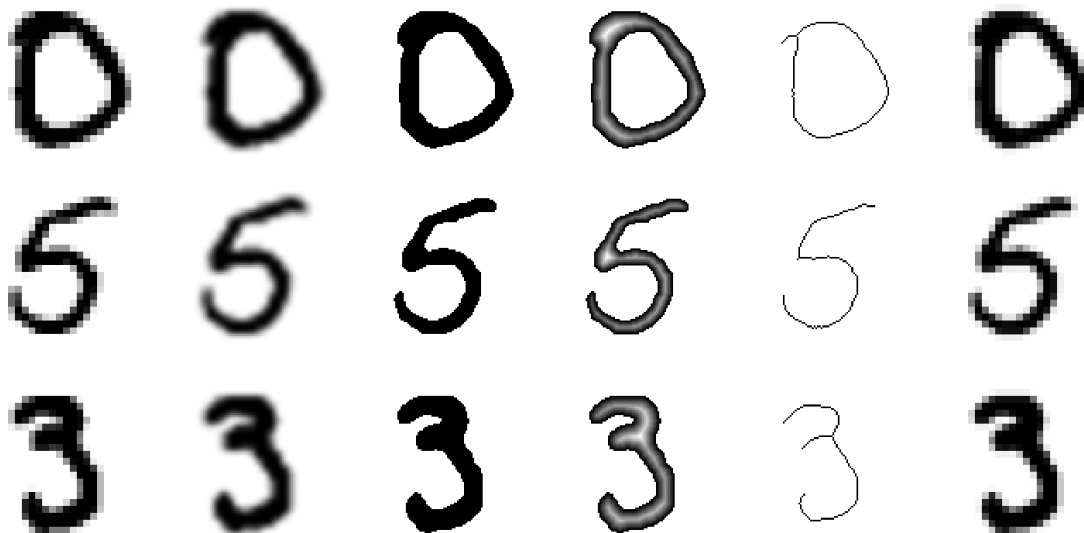
Figure 2: Stages of the image processing pipeline. *Left to right:* original image, upscaled image, binarised image, distance transform, skeleton, downscaled image.

2018). Yet another method, explored in Kumar et al. (2018), is based on the predictive accuracy of individual latent variables to each generative factor. The concurrent work of Eastwood and Williams (2018) introduced a comprehensive methodology for characterising different aspects of representation disentanglement.

2. Morphometry

Meaningful morphometrics are instrumental in characterising distributions of rasterised shapes, such as MNIST digits, and can be useful as additional data for downstream learning tasks. We regard the current choice of morphometrics as a minimal useful collection of attributes that are largely independent and can be estimated robustly. Although further more intricate metrics can be defined, they are likely to be entangled with the ones defined here and to be noisy due to the limited resolution of the original images. We begin this section by describing the image processing pipeline employed for extracting the metrics and for applying perturbations (Section 3), followed by details on the computation of each measurement.

2.1. Processing Pipeline

The original 28×28 resolution of the MNIST images is generally not high enough to enable satisfactory morphological processing: stroke properties (e.g. length, thickness) measured directly on the binarised images would likely be inaccurate and heavily quantised. To mitigate this issue and enable sub-pixel accuracy in the measurements, we propose to use the following processing steps:
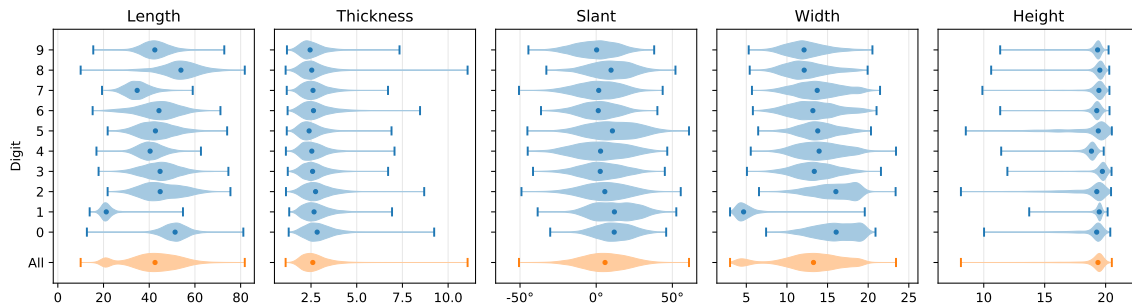
Figure 3: Distribution of morphological attributes in the plain MNIST training dataset

1. *Upscale* (e.g. ×4, to 112×112);[1]

2. *Binarise:* the blurry upscaled image is thresholded at half its intensity range (e.g. 128 for most images), which ensures thin or faint digits are not erased;

3. Compute *Euclidean distance transform* (EDT): each pixel within the digit boundaries contains the distance to its nearest boundary point;

4. *Skeletonise:* detect the ridges of the EDT, i.e. the locus of points equidistant to two or more boundary points (also known as the *medial axis*) (Blum, 1967);

5. Optionally apply *perturbation* to binarised image (cf. Section 3);

6. *Downscale* binarised or perturbed image to original resolution.

   We illustrate the pipeline in Fig. 2. The binary high-resolution digits have smooth boundaries and faithfully capture subtle variations in contour shape and stroke thickness that are only vaguely discernible in the low-resolution images. Additionally, note how the final downscaled image is almost indistinguishable from the original.

   All morphometric attributes described below are calculated for each digit after applying steps 1–4 of this pipeline. The distributions for plain MNIST are plotted in Fig. 3, and the distributions after applying each type of perturbation can be found in Appendix A.

2.2. Stroke Length

Here we approximate the trace of the pen tip, as a digit was being written, by the computed morphological skeleton. In this light, the total length of the skeleton is an estimate of the length of the pen stroke, which in turn is a measure of shape complexity.

   It can be computed in a single pass by accumulating the Euclidean distance of each skeleton pixel to its immediate neighbours, taking care to only count the individual contributions once. This approach is more accurate and more robust against rotations than a naïve estimate by simply counting the pixels.

---

1. Up- and downscaling by a factor of f are done with bicubic interpolation and Gaussian smoothing (bandwidth $\sigma = 2f/6$), following `scikit-image` defaults (van der Walt et al., 2014).
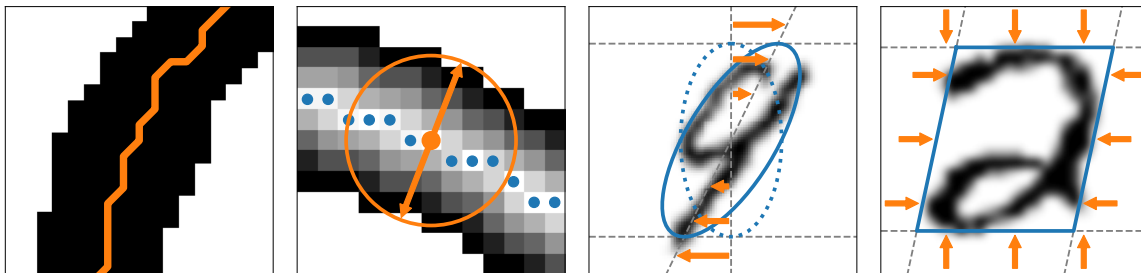
Figure 4: Measuring shape attributes. *Left to right:* stroke length (length of binarised image's skeleton); stroke thickness (average distance transform over skeleton pixels ×2); slant (horizontal shearing angle); width and height (bounding parallelogram).

2.3. Stroke Thickness

A prominent factor of style variation in the MNIST digits is the overall thickness of the strokes, due to both legitimate differences in pen thickness and force applied, and also to the rescaling of the original NIST images by different factors.

We estimate it by exploiting both the computed distance transform and the skeleton. Since skeleton pixels are equidistant to the nearest boundaries and the value of the EDT at those locations corresponds to the local half-width (Blum, 1967), we take twice the mean of the EDT over all skeleton pixels as our global estimate for stroke thickness.

2.4. Slant

The extent by which handwritten symbols lean right or left (forward and backward slant, respectively) is a further notorious and quantifiable dimension of handwriting style. It introduces so much variation in the appearance of characters in images that it is common practice in optical character recognition (OCR) systems to 'deslant' them, in an attempt to reduce within-class variance (LeCun et al., 1998; Teow and Loe, 2002).

We adapt the referred deslanting methodology to describe the slant *angle* of the handwritten digits. After estimating the second-order image moments,

$$\mathbf{S} = \begin{pmatrix} \sum_{i,j} x_{ij}(i - \bar{\imath})^2 & \sum_{i,j} x_{ij}(i - \bar{\imath})(j - \bar{\jmath}) \\ \sum_{i,j} x_{ij}(i - \bar{\imath})(j - \bar{\jmath}) & \sum_{i,j} x_{ij}(j - \bar{\jmath})^2 \end{pmatrix}, \tag{1}$$

we define the slant based on the horizontal shear:

$$\alpha = \arctan\left(-\frac{S_{12}}{S_{22}}\right), \tag{2}$$

where x_{ij} is the intensity of the pixel at row j and column i, and $(\bar{\imath}, \bar{\jmath})$ are the centroid coordinates. The minus sign ensures that positive and negative values of α correspond to forward and backward slant, respectively.

Note that $-\alpha$ is the angle by which the image would need to be horizontally sheared toward the vertical centre line such that the digit is upright (i.e. principal axes aligned with the coordinate axes). Note how this differs from a *rotation* angle, which would instead be defined in terms of $\frac{1}{2}\arctan(2S_{12}/(S_{22} - S_{11}))$.

2.5. Width and Height

It is useful to measure other general shape attributes, such as width, height, and aspect ratio, which also present substantial variation related to personal handwriting style.[2] To this end, we propose to fit a *bounding parallelogram* to each digit, with horizontal and slanted sides (see Fig. 1). We highlight that it does not suffice to simply fit a bounding box with perpendicular sides, since the width would clearly be confounded with slant. These metrics would therefore contain redundant information that obfuscates the true digit width.

The image is swept top-to-bottom with a horizontal boundary to compute a vertical marginal cumulative distribution function (CDF), and likewise left-to-right with a *slanted* boundary for a horizontal marginal CDF, with angle α as computed above. The bounds are then chosen based on equal-tailed intervals containing a given proportion of the image mass—98% in both directions (1% from each side) proved accurate and robust in our experiments.

3. Perturbations

As discussed in Section 1, we bring forward a number of morphological perturbations for MNIST digits, to enable diverse applications and experimentation. In this section, we detail these parametrisable transformations, categorised as global or local.

3.1. Global: Thinning and Thickening

The first pair of transformations we present is based on simple morphological operations: the binarised image of a digit is dilated or eroded with a circular structuring element. Its radius is set proportionally to the estimated stroke thickness (Section 2.3), so that the overall thickness of each digit will decrease or increase by an approximately fixed factor (here, -70% and +100%; see Figs. B.1 and B.2).

Since there is substantial thickness variability in the original MNIST data (cf. Fig. 3) and most thinned and thickened digits look very plausible, we believe that these perturbations can constitute a powerful form of data augmentation for training. For the same reason, we have not included these perturbations in the abnormality detection experiments (Section 4.3.3).

3.2. Local: Swelling

In addition to the global transformations above, we introduce *local* perturbations with variable location and extent, which are harder to detect automatically. Given a radius R, a centre location $\mathbf{r}_0$ and a strength parameter $\gamma > 1$, the coordinates $\mathbf{r}$ of pixels within distance R of $\mathbf{r}_0$ are nonlinearly warped according to a radial power transform:

$$\mathbf{r} \mapsto \mathbf{r}_0 + (\mathbf{r} - \mathbf{r}_0)\Big(\frac{\|\mathbf{r} - \mathbf{r}_0\|}{R}\Big)^{\gamma-1}, \tag{3}$$

leaving the remaining portions of the image untouched and resampling with bicubic interpolation.

---

2. Little variation in height is expected, since the original handwritten digits were scaled to fit a 20×20 box (LeCun et al., 1998). Nevertheless, a minority of digits were originally wider than they were tall, which explains the long tails in the distribution of heights (Fig. 3).

In the experiments and released dataset, we set $\gamma = 7$ and $R = 3\sqrt{\theta}/2$, where θ is thickness. Unlike simple linear scaling with θ, this choice for R produces noticeable but not exaggerated effects across the thickness range observed in the dataset (cf. Fig. B.3). The centre location, $\mathbf{r}_0$, is picked uniformly at random from the pixels along the estimated skeleton.

3.3. Local: Fractures

Finally, we describe the proposed procedure for adding fractures to an MNIST digit, where we define a fracture as a break in the continuity of a pen stroke. Because single fractures can in many cases be easily mistaken for true gaps between strokes, we add multiple fractures to each affected digit.

When selecting where to place a fracture, we attempt to avoid locations too close to stroke tips (points on the skeleton with a single neighbour) or fork points (more than two neighbours). This is achieved by sampling only among those skeleton pixels above a certain distance to these detected points. In addition, we desire fractures to be transversal to the pen strokes. Local orientation is determined based on second-order moments of the skeleton inside a window centred at the chosen location, and the length of the fracture is estimated from the boundary EDT. Finally, the fracture is drawn onto the high-resolution binary image with a circular brush along the estimated normal.

In practice, we found that adding three fractures with $1.5\,\mathrm{px}$ thickness, $2\,\mathrm{px}$ minimum distance to tips and forks and angle window of $5{\times}5\,\mathrm{px}^2$ ('px' as measured in the low resolution image) produces detectable but not too obvious perturbations (see Fig. B.4). We also extend the lines on both ends by $0.5\,\mathrm{px}$ to ensure some tolerance.

4. Evaluation Case Studies

In this section, we demonstrate potential uses of the proposed framework: using morphometrics to characterise the distribution of samples from generative models and finding associations between learned latent representations and morphometric attributes. In addition, we exemplify in Section 4.3 a variety of novel ways to make use of morphometry and perturbations for evaluating predictive models on MNIST and beyond.

4.1. Sample Diversity

Here we aim to illustrate ways in which the proposed MNIST morphometrics may be used to visualise distributions learned by generative models and to quantify their agreement with the true data distribution in terms of these concrete attributes. Moreover, extracting such measurements from model samples could be a step toward diagnosing mode collapse.

We exemplify this scenario with a generative adversarial network (GAN, Goodfellow et al., 2014) and a β-variational autoencoder (β-VAE, Higgins et al., 2017), both with generator (resp. decoder) and discriminator architecture as used in the MNIST experiments in Chen et al. (2016), and encoder mirroring the decoder. We train a β-VAE with $\beta = 4$, as in Higgins et al. (2017)'s experiments on small binary images, and a vanilla GAN with non-saturating loss, both with 64-dimensional latent space. To explore the behaviour of a much less expressive model, we additionally train a GAN with only two latent dimensions.

(a) MNIST test images

(b) Samples from a β-VAE

(c) Samples from a GAN
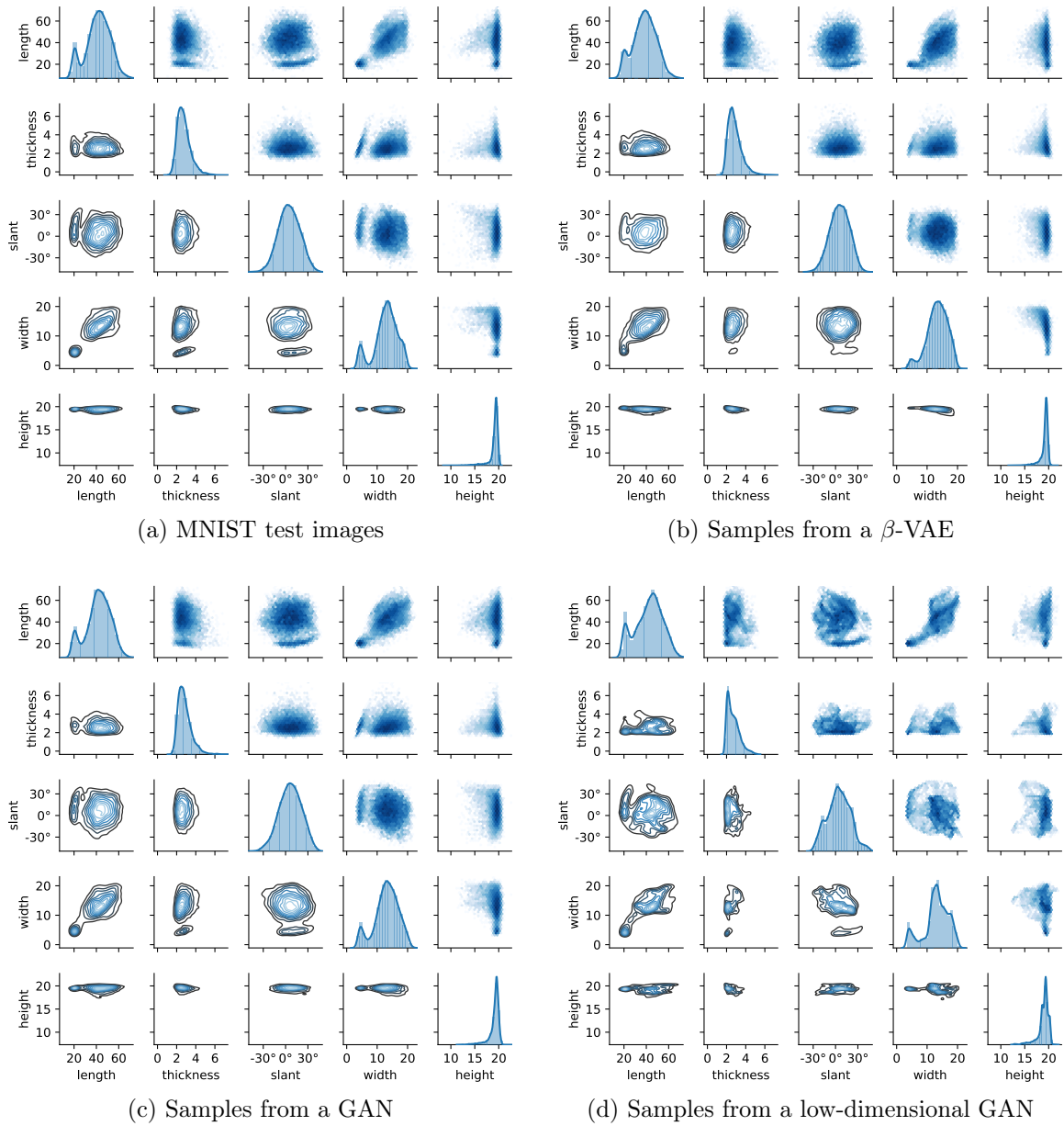
(d) Samples from a low-dimensional GAN

Figure 5: Distribution of morphometric attributes for MNIST test dataset and samples from some generative models. Diagonals show marginal histograms and KDEs, upper-triangular plots show pairwise log-histograms and lower-triangular plots show pairwise KDEs.

Test data vs.	Dims.	$\mathrm{MMD}_l^2 \pm$ std. error ($\times 10^{-3}$)	p
β-VAE	64	0.792 ± 1.569	.3068
GAN	64	1.458 ± 1.650	.1885
GAN	2	**8.876 ± 1.807**	**.0000**

Table 1: Kernel two-sample tests between model samples and true test data. The bold row indicates a significant divergence, thus a failure of the model in faithfully reproducing the data distribution.

4.1.1. VISUALISATION

Figure 5 illustrates the morphometric distributions of the plain MNIST test images and of 10,000 samples from each of these three models. As can be seen, morphometrics provide *interpretable* low-dimensional statistics which allow comparing distributions learned by generative models between each other and with true datasets. While Figs. 5b and 5c show model samples roughly as diverse as the true images, the samples from the low-dimensional GAN in Fig. 5d seem concentrated on certain regions, covering a distribution that is less faithful to the true one in Fig. 5a.

4.1.2. STATISTICAL COMPARISON

We argue that in this lower-dimensional space of morphometrics it is possible to statistically compare the distributions, since this was shown not to be effective directly in image space (e.g. Theis et al., 2016). To this end, we propose to use kernel two-sample tests based on maximum mean discrepancy (MMD) between morphometrics of the test data and of each of the sample distributions (Gretton et al., 2012). The MMD statistic is a measure of dissimilarity between distributions, such that a value significantly greater than zero (small p-value) suggests a detectable difference. Here, we performed the linear-time asymptotic test described in Gretton et al. (2012, §6) (details and further considerations in Appendix C).

The results of these tests are displayed in Table 1, where we observe that samples from the high-dimensional β-VAE and GAN did not exhibit a significantly large MMD, thus there is no evidence of distributional mismatch. On the other hand, the low-dimensional GAN's samples show a significant departure from the data distribution, confirming the qualitative judgement based on comparing Figs. 5a and 5d.

4.2. Disentanglement

In this experiment, we demonstrate that: (a) standard MNIST can be augmented with morphometric attributes to quantitatively study representations computed by an *inference* model (as already possible with annotated datasets, e.g. dSprites and 3D faces); (b) we can measure shape attributes of samples to assess disentanglement of a *generative* model, which is unprecedented to the best of our knowledge; and (c) this analysis can also diagnose when a model unexpectedly *fails* to disentangle a known aspect of the data.

As for the sample diversity experiments in Section 4.1, the purpose of this disentanglement analysis is to study the behaviour of specific trained model instances, and results may not

generalise to the respective model classes or architectures, to different datasets, or even to distinct runs of the same set-up. Similarly to any other evaluation methodology, findings concerning disentanglement patterns are reproducible only to the extent that the model under study can be reliably retrained to a similar state.

4.2.1. Methodology

Using an approach related to the disentanglement measure introduced by Kumar et al. (2018), we study the correlation structures between known generative factors and learned latent codes. Specifically, we compute the *partial correlation* $(r_{c_i y_j \cdot \mathbf{c}_{-i}})$ between each latent code variable (c_i) and each morphometric attribute (y_j), controlling for the variation in the remaining latent variables $(\mathbf{c}_{-i})$.[3] As opposed to the simple correlation, this technique allows us to study the *net* first-order effect of each latent code, all else being equal.

This partial correlation is calculated as follows, ranging over codes i and attributes j:

$$r_{c_i y_j \cdot \mathbf{c}_{-i}} = -\frac{P^{(j)}_{1,i+1}}{\sqrt{P^{(j)}_{11} P^{(j)}_{i+1,i+1}}} \quad \in [-1, 1], \tag{4}$$

where $\mathbf{P}^{(j)}$ is the sample precision matrix (i.e. inverse of sample covariance) of $(y_j, c_1, \ldots, c_D)$ over all data points. Note that Eq. (4) is closely related to the least-squares coefficients for multiple linear regression of y_j on $(c_1, \ldots, c_D)$, and its magnitude can be directly interpreted as a relative importance score within the framework of Eastwood and Williams (2018).

We plot the calculated partial correlations in the form of tables, where each row corresponds to a shape attribute, and columns represent the latent variables. These tables should be read row-wise, as the scores are normalised separately for each shape factor. Entries close to zero do not indicate that the latent dimension is unrelated to the measured attribute, but rather that it is entangled with others and has little marginal effect on its own.

In addition, we demonstrate how our morphometry framework is fully compatible with contemporary disentanglement metrics by also computing the mutual information gap (MIG score; Chen et al., 2018) in each scenario. This score is based on the mutual information of each pair of generative factor (here, shape attribute) and latent code dimension, and reports the gap between the two leading codes for each factor, normalised to $[0, 1]$. It therefore quantifies how prominently the attribute is represented by a single latent dimension, i.e. its *completeness* in the terminology of Eastwood and Williams (2018). The overall MIG for each model is given by the average over all attributes.

For these experiments, we trained InfoGAN (Chen et al., 2016) instances with various latent space configurations, then took MAP estimates of latent codes for each image (i.e. maximal logit for categorical codes and mean for continuous codes) as predicted by the variational recognition network. Models were trained for 20 epochs using 64 images per batch, with no hyperparameter tuning. We emphasise that our goal was to illustrate how the proposed morphometrics can serve as tools to better understand whether models behave as intended, and not to optimally train them in each scenario.

---

3. For the categorical code, c_1, we take a single binary dummy variable for each category, $c_1^{(k)}$, while controlling only for the remaining codes (c_2, c_3 etc.) to avoid multicollinearity.

	# Cat.	# Cont.	# Bin.	Dataset	
InfoGAN-A	10	2	0	plain:	plain only
InfoGAN-B	10	3	0	global:	plain + thinning + thickening
InfoGAN-C	10	2	2	local:	plain + swelling + fractures

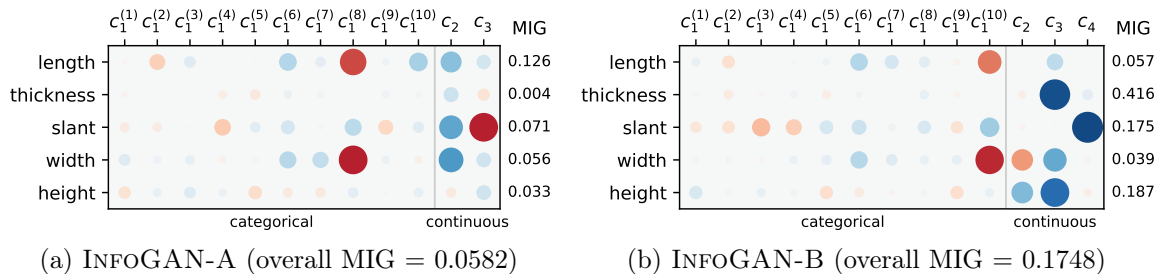Table 2: Settings for InfoGAN disentanglement experiments

(a) InfoGAN-A (overall MIG = 0.0582) (b) InfoGAN-B (overall MIG = 0.1748)

Figure 6: Partial correlations between inferred latent codes and morphometrics of test images. Circle area and colour strength are proportional to correlation magnitude, blue is positive and red is negative. On the right, we indicate the mutual information gap score (MIG) for each attribute.

4.2.2. Inferential Disentanglement

To illustrate how this methodology can be applied in practice to assess disentanglement, we consider two settings. The first is the same as in the MNIST experiment from Chen et al. (2016), with a 10-way categorical and two continuous latent codes, trained and evaluated on the plain MNIST digits, to which we will refer as InfoGAN-A.

The second setting was designed to investigate whether the model could disentangle the concept of *thickness*, by including an additional continuous latent code and training on a dataset with exaggerated thickness variations. We constructed this dataset by randomly interleaving plain, thinned and thickened digit images in equal proportions. Since the perturbations were applied completely at random, we expect a trained generative model to identify that thickness should be largely independent of the other morphological attributes. We refer to this set-up as InfoGAN-B. Table 2 summarises the different experimental settings, for reference.

In Fig. 6a, we see that InfoGAN-A learned to encode slant mostly in c_3, while $c_1^{(8)}$ clearly relates to the '1' class (much narrower digit shape and shorter pen stroke; cf. Fig. 3). Figure 6b quantitatively confirms the hypothesis that InfoGAN-B's recognition network would learn to separate slant and thickness (in c_4 and c_3, resp.), the most prominent factors of style variation in this dataset. Interestingly, it shows that c_3 also associates with height, as thicker digits also tend to be taller.

Moreover, we observe that the MIG for each attribute correlates with the qualitative interpretation of these partial correlation tables. For example, thickness is represented considerably more compactly in InfoGAN-B (mainly by c_3) than in InfoGAN-A (highly entangled), exhibiting a correspondingly much higher MIG score. For InfoGAN-B, a

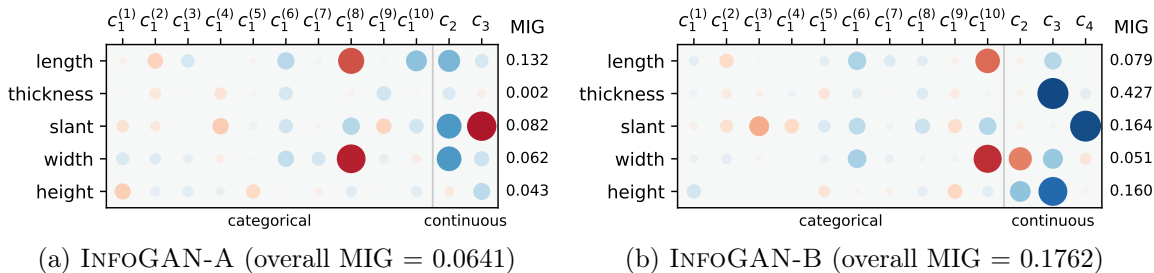(a) INFOGAN-A (overall MIG = 0.0641)   (b) INFOGAN-B (overall MIG = 0.1762)

Figure 7: Partial correlations between 1000 sampled latent codes and morphometrics of the corresponding generated images

noticeable portion of the slant variation was additionally captured by c_1, which is reflected in its lower MIG than that for thickness. Lastly, the difference in overall MIG between models validates our qualitative judgement about global disentanglement.

4.2.3. GENERATIVE DISENTANGLEMENT

The evaluation methodology described above is useful for investigating the behaviour of the *inference* direction of a model, and can readily be used with datasets which include ground-truth generative factor annotations. On the other hand, unless we trust that the inference approximation is highly accurate, this tells us little about the *generative* expressiveness of the model. This is where computed metrics truly show their potential: one can measure generated samples, and investigate how their attributes relate to the latent variables used to create them.

Figure 7 shows results for a similar analysis to Fig. 6, but now evaluated on *samples* from that model. As the tables are mostly indistinguishable, we may argue that in this case the inference and generator networks have learned to consistently encode and decode the digit shape attributes. Further, the per-attribute and overall MIG scores agree once more with the observed disentanglement patterns.

As further illustration, Fig. 8 displays traversals of the latent space, obtained by varying a subset of the latent variables while holding the remaining ones (including noise) constant. With these examples, we can qualitatively verify the quantitative results in Fig. 7. Note that, until now, visual inspection was typically the only means of evaluating disentanglement and expressiveness of the generative direction of image models (Kim and Mnih, 2018).

4.2.4. DIAGNOSING FAILURE

We also attempted to detect whether an InfoGAN had learned to discover that local perturbations (swelling and fractures) correspond to independent generative factors. To this end, we extended the model formulation with additional Bernoulli latent codes, which may learn to encode presence/absence of each type of local perturbation. The model investigated here, dubbed INFOGAN-C (cf. Table 2), had a 10-way categorical, two continuous and two binary codes, and was trained with a dataset of plain, swollen and fractured digits (randomly mixed as above).

15

(a) InfoGAN-A: one-dimensional traversals of c_1 (*top*, 'digit type') and c_3 (*bottom*, 'slant'). Samples in each row share the values of remaining latent variables and noise.

(b) InfoGAN-B: two-dimensional traversal of $c_4 \times c_3$ ('thickness' $\times$ 'slant')
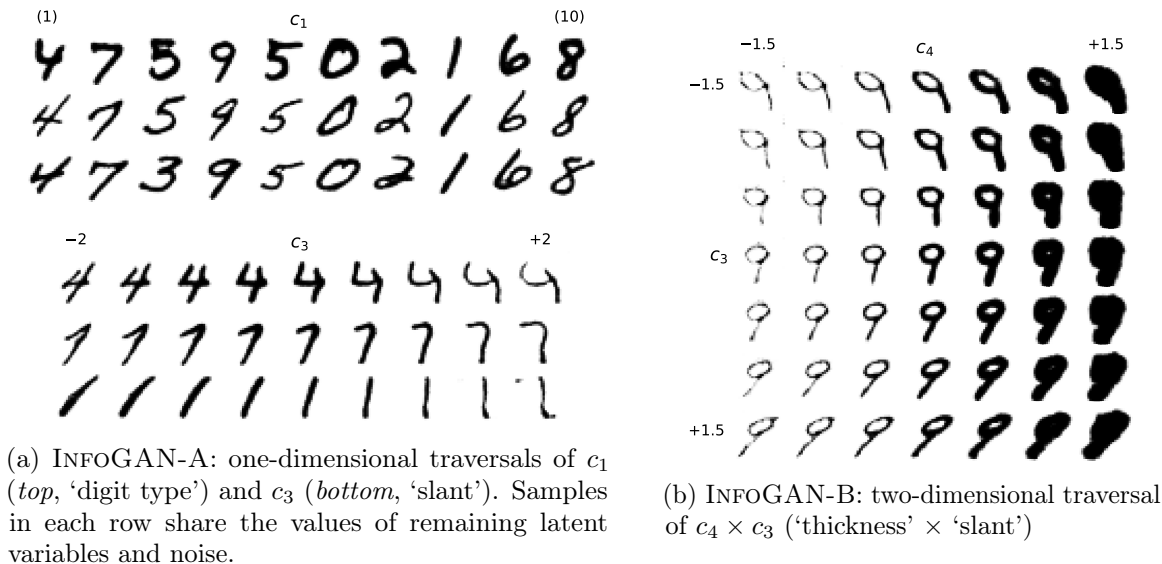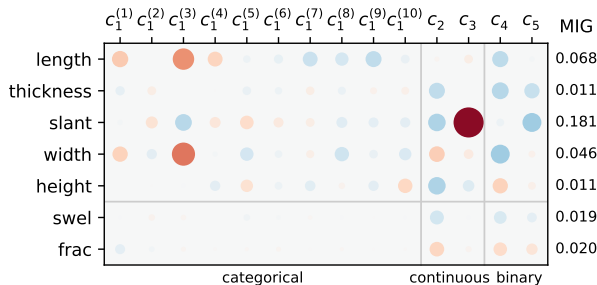
Figure 8: InfoGAN latent space traversals

Figure 9: Partial correlations between inferred latent codes and morphometrics of test images for InfoGAN-C (overall MIG = 0.0507). 'swel' and 'frac' refer to the binary perturbation labels.

Again via inferential partial correlation analysis—now including ground-truth perturbation labels—we can quantitatively verify that this particular model instance was unable to meaningfully disentangle the perturbations (Fig. 9, bottom-right block). This is also confirmed by the low MIG for the corresponding rows. In fact, it appears that the addition of the binary variables did not lead to more expressive representations in this case, even impairing the disentanglement of the categorical variables, if compared to Figs. 6a and 6b, for example.

4.3. Predictive Model Evaluation

Although the driving motivation for introducing Morpho-MNIST has been the lack of means for quantitative evaluation of generative models, the proposed framework may also be a valuable resource in the context of predictive models. In this section we present several case studies to demonstrate how the proposed perturbations can be useful for systematically

Test data	kNN	SVM	MLP	CNN
PLAIN	96.25	95.71	97.97	98.95
LOCAL	95.22	92.47	93.15	95.33

Table 3: Example results on digit recognition (accuracy, %) under synthetic domain shift using the proposed data perturbations

evaluating predictive robustness under controlled domain shift and how morphometry enables an entirely new variety of experimental settings and nuanced analyses based on shape stratification. Further we demonstrate how the perturbation labels and morphometrics can serve as prediction targets in their own right for new supervised tasks. For simplicity and reproducibility, we use the same datasets as in the disentanglement experiments (Section 4.2): plain digits (PLAIN), plain mixed with thinned and thickened digits (GLOBAL), and plain mixed with swollen and fractured digits (LOCAL).

4.3.1. ROBUSTNESS

Synthetic perturbations enable fabrication of domain shift scenarios, aiming to systematically evaluate the robustness of predictive models to specific types of variation. For example, digits with local perturbations (swelling and fractures) are still clearly identifiable by humans (see Figs. B.3 and B.4). However, a practitioner might be interested in investigating how different types of classifiers react to these changes.

Here we compare simple representative methods from four classification paradigms: neighbourhood-based, kernel-based, fully connected, and convolutional. Specifically, we evaluated k-nearest-neighbours (kNN) using $k = 5$ neighbours and ℓ_1 distance weighting, a support vector machine (SVM) with polynomial kernel and penalty parameter $C = 100$, a multi-layer perceptron (MLP) with 784–200–200–L architecture (L: number of outputs), and a LeNet-5 (LeCun et al., 1998) convolutional neural network (CNN).

Each model is trained for digit recognition once on PLAIN, then tested separately on both PLAIN and LOCAL test datasets, to investigate the effect of *domain shift*. All methods suffer a drop in test accuracy on LOCAL (Table 3), and, considering that one third of the images in LOCAL are identical to ones in PLAIN, the absolute performance drops for the perturbed digits alone are in fact in the order of 1.5%, 4.9%, 7.2%, and 5.4%, respectively. kNN appears to be the most robust to local perturbations, conceivably because they affect only a few pixels, leaving the image distance between neighbours largely unchanged. Contrarily, both the global patterns that SVM and MLP rely on and the CNN's local filter responses may have changed considerably. Such analyses can provide new insights into the behaviour, strengths, and weaknesses of different models, and can be useful to inform new strategies for developing more robust approaches.

4.3.2. DATA STRATIFICATION

Originally, MNIST could be stratified exclusively by digit class, as this was the only meta-information available. However, with morphometrics, data can be sliced and grouped

(a) Uncertainty of MC-dropout model

(b) Auto-encoder reconstruction error (left, ●) and thickness of reconstructed digits (right, ▲)
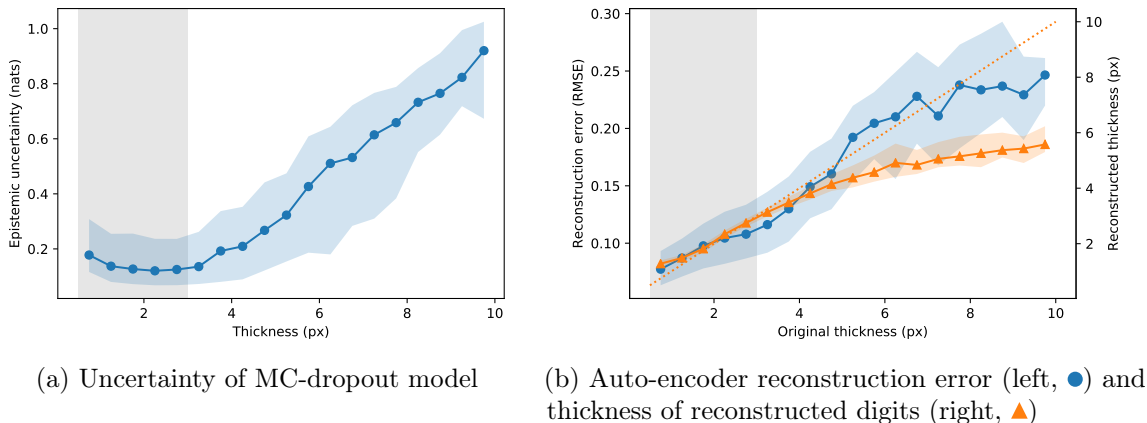
Figure 10: Results for toy outlier detection experiments showcasing data stratification by stroke thickness: training digits were restricted to those thinner than 3 px (shaded range), and test digits were grouped into 0.5 px-wide bins for analysis. Lines and shaded regions indicate medians and interquartile ranges.

according to entirely new criteria: we are able to select the shape properties we desire in the training set and also perform finer analyses at test time. It is even possible to combine multiple attributes for multifaceted studies.

To illustrate how these capabilities can be put to use, we present a simplified case study on a hypothetical unsupervised outlier detection scenario. We take the GLOBAL dataset, restrict the training data to digits whose thickness is below 3 px, and evaluate the model's response to test digits across the entire thickness range, especially out-of-distribution samples. Here we assume the true selection criterion is unknown at learning time. Two example frameworks were considered: uncertainty-based and reconstruction-based.

For the first setting, we assume a research question about outlier detection with respect to the digit classification task. We take a LeNet-5 classifier (LeCun et al., 1998) and employ Monte Carlo dropout (Kendall and Gal, 2017) to estimate the model's epistemic uncertainty via mutual information (Smith and Gal, 2018). This metric, which quantifies the model's uncertainty about its parameters when presented with a test input, has been shown to be effective at detecting out-of-distribution samples (Smith and Gal, 2018). Figure 10a shows that the model seems markedly more uncertain on average regarding thicker digits (e.g. above 5 px), although there is no abrupt jump in uncertainty at the inlier boundary. Since we have 'ground-truth' outlier labels for the test images based on their thickness, we are also able to evaluate the discriminative power of this score: with an area under ROC curve (AUC) of only 0.67, this approach did not result in a particularly strong outlier detector.

In the second setting, we explored outlier detection with a deep convolutional auto-encoder.[4] It was likewise trained on digits thinner than 3 px, and we plot in Fig. 10b its reconstruction error for test digits of varying thicknesses. We observe a steeper trend than

---

4. The encoder has three convolutional and one fully connected layers, and the mirrored decoder employs transposed convolutions. The convolutional layers have 3×3 filters with stride 2 and 32–64–128 feature channels, and the bottleneck layer is 10 units wide.

Prediction task	kNN	SVM	MLP	CNN
Abnormality detection (%)	65.10	77.59	88.25	97.53
Thickness regression (RMSE)	0.467	0.365	0.348	0.279

Table 4: Example results on some of the new MNIST prediction tasks

for the previous model's uncertainty (Fig. 10a), and in fact this auto-encoder's reconstruction error is far more discriminative (AUC = 0.85). In addition, it is also possible to collect morphometrics of the reconstructions and visualise to what extent the original shape attributes are preserved. Figure 10b suggests the interesting finding that the reconstructions of thick digits are consistently thinner than the originals, as if the auto-encoder 'projected' these digits towards the training subspace to some extent.

4.3.3. New Prediction Tasks

In the previous two sections, we demonstrated how perturbations may be leveraged to systematically evaluate robustness to domain shift and how morphometry can prove valuable for selecting training data and for performing stratified test-time analysis. Here we illustrate how perturbation labels and morphometric attributes can be used for interesting new prediction tasks on MNIST, evaluating the same models from Section 4.3.1.

As our first example, we define a supervised abnormality detection task, using the LOCAL dataset, to predict whether a digit is normal or perturbed (swollen or fractured)—compare with lesion detection in medical scans. The top row of Table 4 indicates that the CNN was able to detect abnormalities with high accuracy, likely thanks to local invariances of its architecture, whereas the shallow classifiers, kNN and SVM, performed remarkably worse. Note how the performance is generally much lower than for digit classification, revealing the higher difficulty of this binary problem compared to the ten-class recognition task.

Finally, we also constructed a regression task for digit thickness using the GLOBAL dataset, mimicking medical imaging tasks such as estimating brain age from cortical grey matter maps. Since this is a non-trivial task requiring some awareness of local geometry, it is unsurprising that the convolutional model outperformed the others, which rely on holistic features (Table 4, bottom row).

5. Discussion and Perspectives

Through concrete examples, we demonstrated in Section 4 how morphometry and perturbations can be exploited 1) for qualitative and, more importantly, *quantitative* evaluation of generative models in terms of sample diversity and disentanglement, and 2) in developing novel experimental protocols for analysing predictive models. The presented methodologies are broadly applicable to datasets of rasterised shapes beyond MNIST.

For example, we apply our exact same morphometry pipeline to handwritten digits from the USPS dataset (LeCun et al., 1990), also composed of small grayscale images (16×16). Example images from this dataset and distributions of length, thickness, width, and height can be seen in Fig. 11. Slant was excluded because the USPS digits were already deslanted, with residual variation in the order of ±1°. Another popular dataset to which our morphometrics

(a) USPS digits

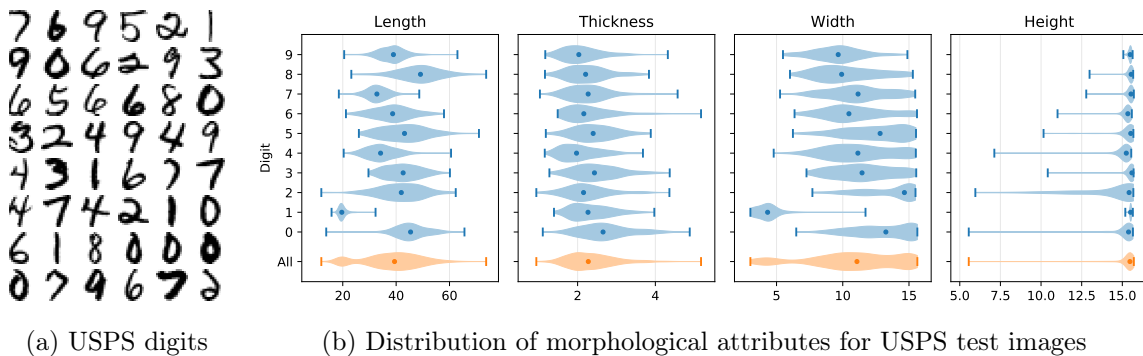(b) Distribution of morphological attributes for USPS test images

Figure 11: Application of the proposed morphometrics to the USPS dataset

can be directly applied is Omniglot (Lake et al., 2015), consisting of handwritten symbols from 50 different alphabets.

Besides, the general perspective of measuring attributes directly from images is evidently not limited to datasets of handwritten characters. Consider for example the dSprites dataset (Matthey et al., 2017), often used in disentanglement research, which contains 64×64 synthetic binary images of small shapes with ground-truth annotations. Attributes such as location and scale can easily be measured based on image moments (see Section 2.4). They are expected to be robust even for blurry generated samples. As discussed previously, while the annotations provided with such data facilitate evaluation of inferential disentanglement, these measurements enable the unprecedented analysis of generative behaviour.

An obvious limitation of morphometry is the tacit assumption that generated samples resemble true data closely enough that the shape measurements are indeed meaningful. What happens if the generative model under study is underfitted and produces nonsensical images? In such case, one expects the joint distribution of sample morphometrics to be strikingly different from that of true data, even if some individual metrics may have been extracted successfully (e.g. width and height of a reasonably sized blur). Like mode collapse, this issue can be diagnosed straightforwardly—for example with an analysis similar to that from Section 4.1—and may constitute an interesting finding by itself.

6. Conclusion

With Morpho-MNIST we provide a number of mechanisms to quantitatively assess representation learning with respect to measurable factors of variation in the data. We believe that this may prove an important asset for future research on generative models, and will hopefully also be useful in evaluating predictive models. We emphasise that the proposed morphometrics can be used *post hoc* to evaluate already trained models, potentially revealing novel insights about their behaviour. In addition, we have shown that our morphometry and perturbation framework easily extends to datasets other than MNIST. We hope that the case studies presented here may also invite research into developing metrics for other kinds of images beyond rasterised 2D shapes or even to non-imaging data, paving the way to more widespread analysis of true generative performance.

Acknowledgments

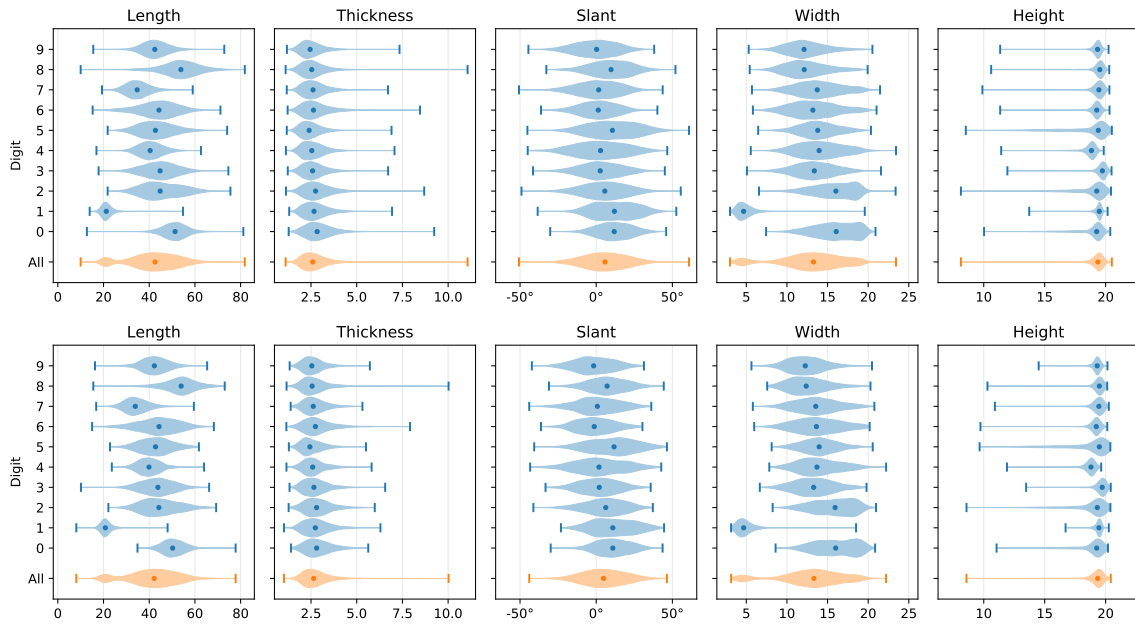## Appendix A. Morphometrics of Plain and Perturbed Datasets

Figure A.1: Distribution of morphological attributes for plain MNIST digits. *Top:* training set; *bottom:* test set.
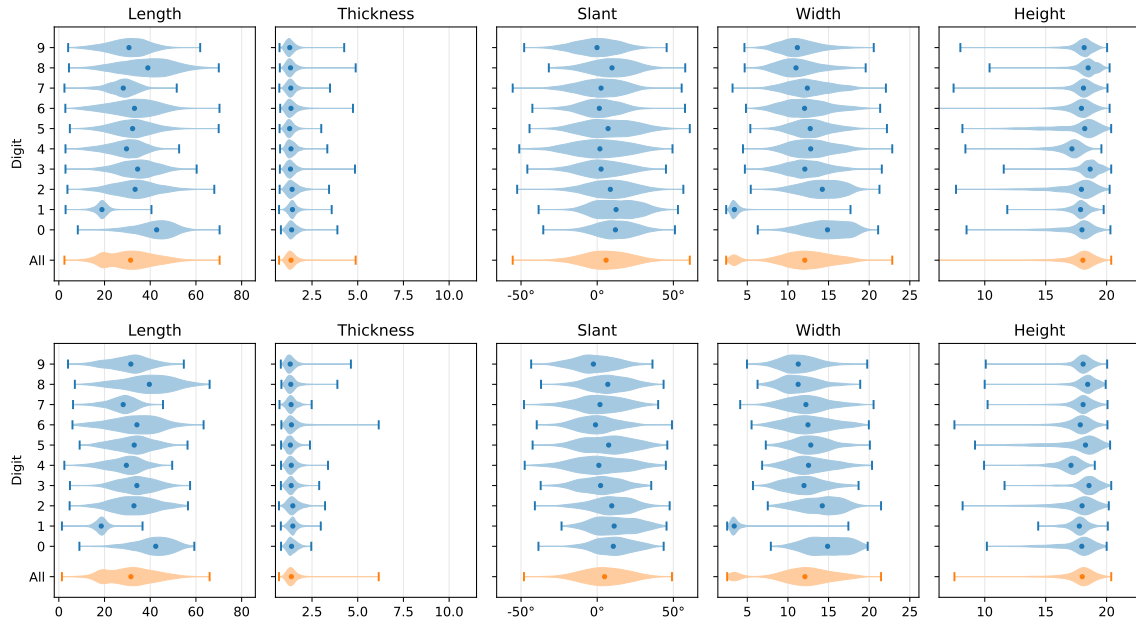
Figure A.2: Distribution of morphological attributes for thinned MNIST digits. *Top:* training set; *bottom:* test set.
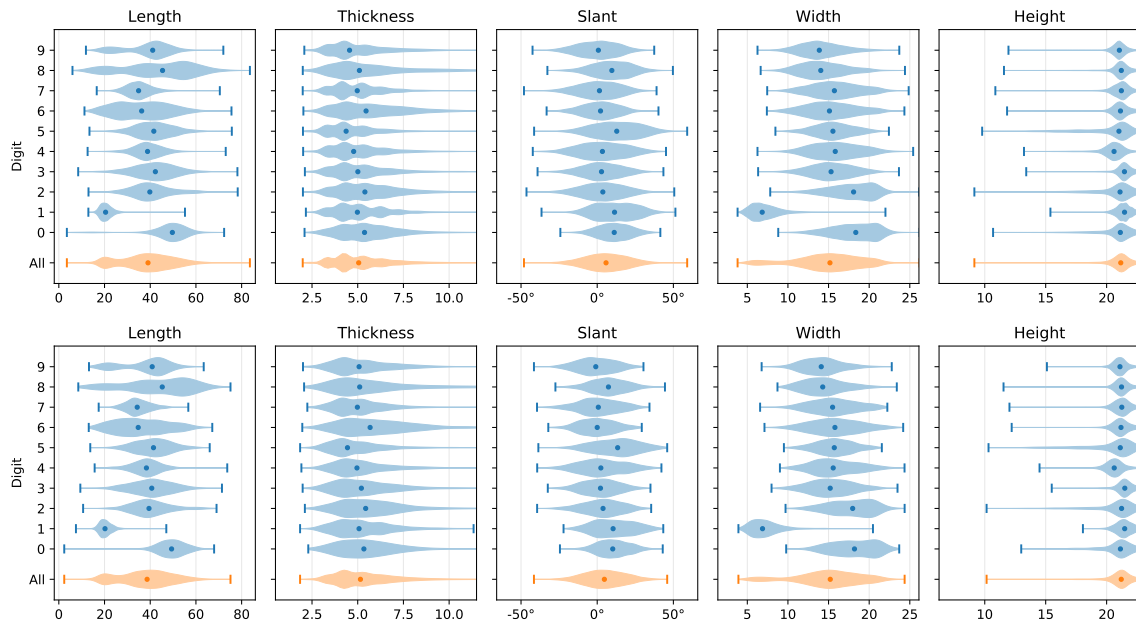

Figure A.3: Distribution of morphological attributes for thickened MNIST digits. *Top:* training set; *bottom:* test set.
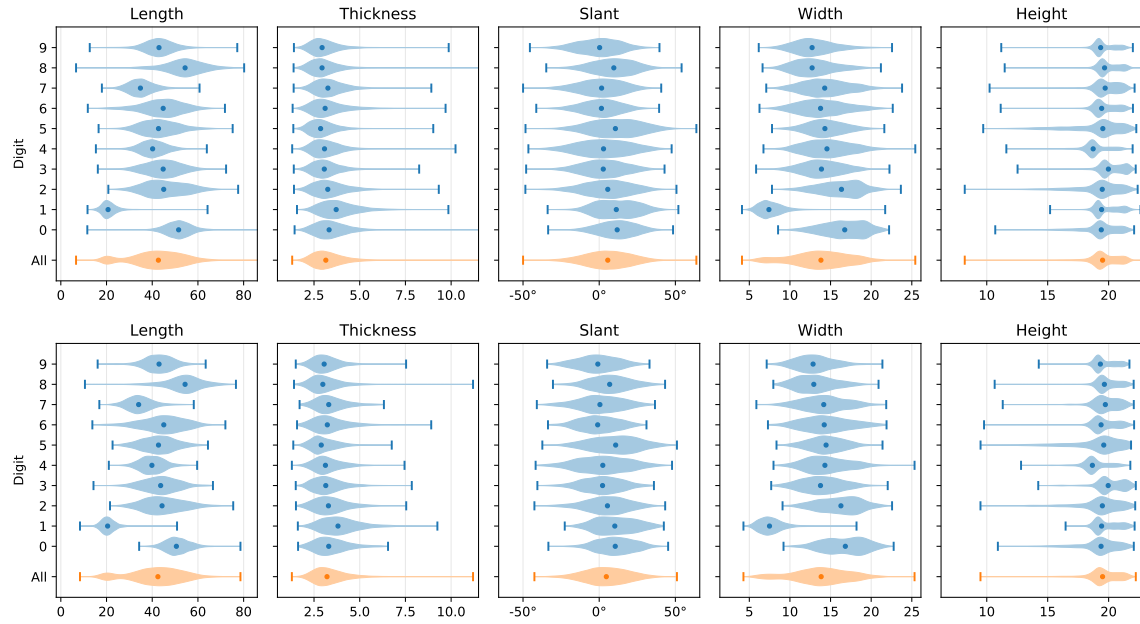
Figure A.4: Distribution of morphological attributes for swollen MNIST digits. *Top:* training set; *bottom:* test set.
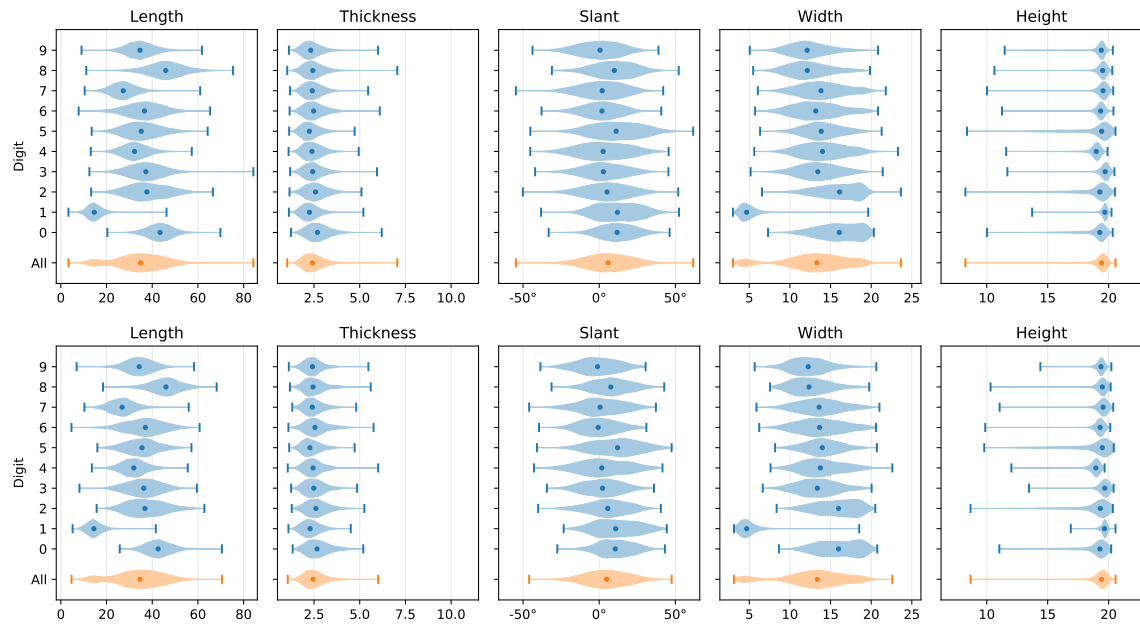

Figure A.5: Distribution of morphological attributes for fractured MNIST digits. *Top:* training set; *bottom:* test set.

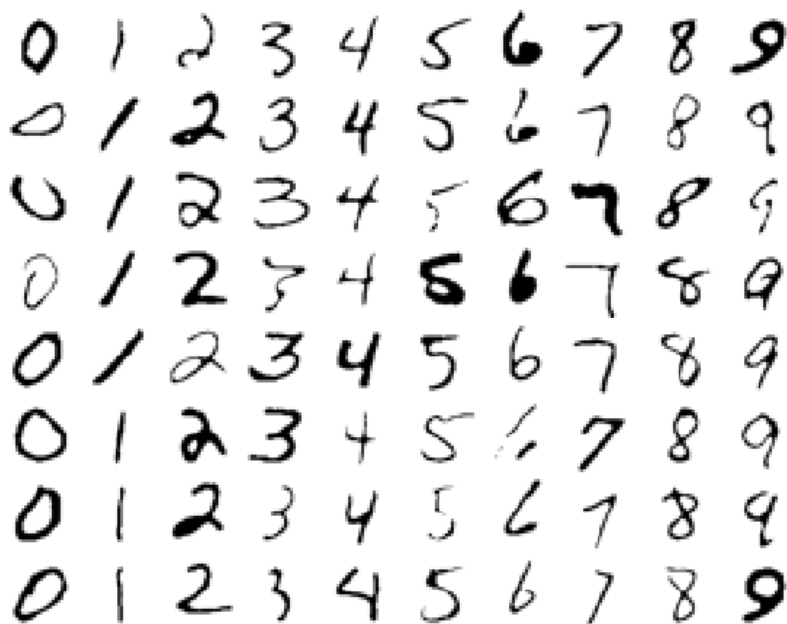# Appendix B. Perturbation Examples

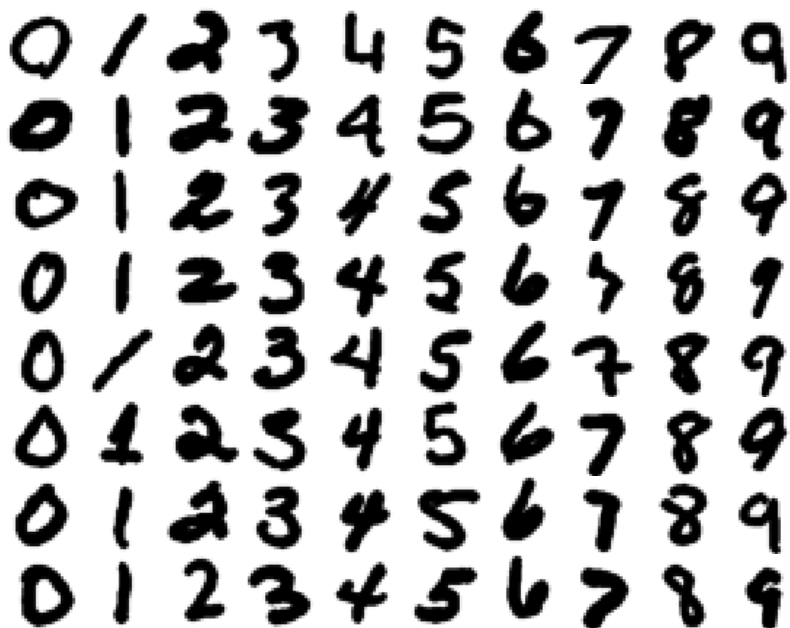Figure B.1: Examples of globally thinned digits

Figure B.2: Examples of globally thickened digits

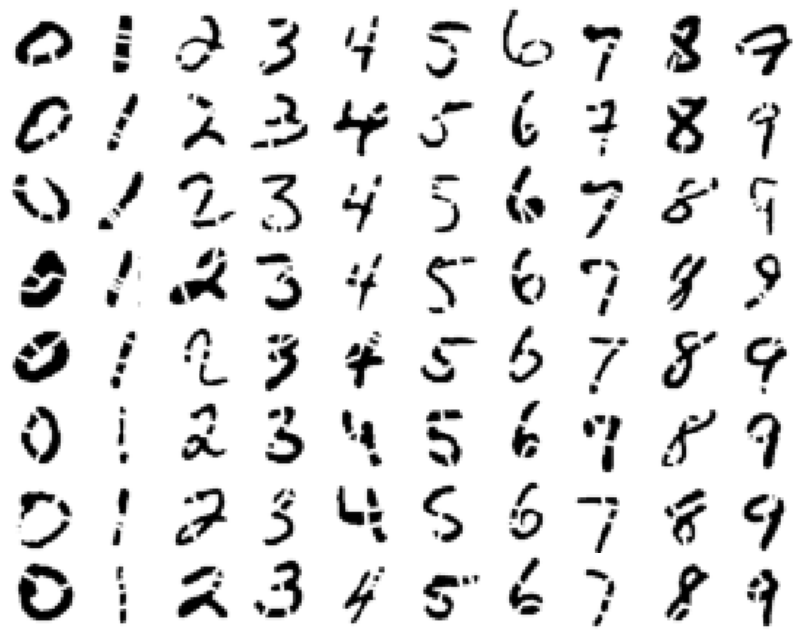Figure B.3: Examples of digits with local swelling

Figure B.4: Examples of digits with local fractures

Appendix C. MMD Details

We employed a Gaussian product kernel with bandwidths derived from Scott's rule, analogously to the KDE plots in Fig. 5. Scott's rule of thumb defines the bandwidth for a density estimation kernel as $N^{-1/(D+4)}$ times the standard deviation in each dimension, where N and D denote sample size and number of dimensions (Scott, 1992, Eq. 6.42). We determine the KDE bandwidths separately for real and sample data, then add their squares to obtain the squared bandwidth of the MMD's Gaussian kernel, as it corresponds to the *convolution* of the density estimation kernels chosen for each set of data. See Gretton et al. (2012, §3.3.1) for further details on the relation between MMD and L_2 distance of kernel density estimates.

Whereas the bandwidth heuristic used here is fairly crude, much more sophisticated kernel selection procedures exist, e.g. by explicitly optimising the test power (Sutherland et al., 2017). A further analysis tool in a similar vein would be to apply a *relative* MMD similarity test (Bounliphone et al., 2016), to rank trained models based on sample fidelity. It would also be possible to adopt a model criticism methodology based on the MMD witness function (Lloyd and Ghahramani, 2015), to identify over- and under-represented regions in morphometric space (and corresponding generated image exemplars could also be inspected).

References

Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs learn the distribution? Some theory and empirics. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

Mathieu Aubry, Daniel Maturana, Alexei A. Efros, Bryan C. Russell, and Josef Sivic. Seeing 3D chairs: exemplar part-based 2D–3D alignment using a large dataset of CAD models. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, pages 3762–3769. IEEE, 2014. Dataset URL `https://www.di.ens.fr/willow/research/seeing3Dchairs/`.

Mikołaj Bińkowski, Dougal Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

Harry Blum. A transformation for extracting new descriptors of shape. In *Models for the perception of speech and visual form*, pages 362–380. MIT Press, 1967.

Wacha Bounliphone, Eugene Belilovsky, Matthew B. Blaschko, Ioannis Antonoglou, and Arthur Gretton. A test of relative similarity for model selection in generative models. In *International Conference on Learning Representations (ICLR 2016)*, 2016.

Yaroslav Bulatov. notMNIST dataset, sep 2011. URL `https://yaroslavvb.blogspot.co.uk/2011/09/notmnist-dataset.html`.

Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 2610–2620, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 2172–2180, 2016.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters, 2017. arXiv:1702.05373. Dataset URL `https://www.nist.gov/itl/iad/image-group/emnist-dataset`.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 708–718, 2018.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 2672–2680, 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Patrick J. Grother and Kayee K. Hanaoka. NIST Special Database 19. Technical report, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2016. Dataset URL `https://www.nist.gov/srd/nist-special-database-19`.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 6626–6637, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR 2017)*, 2017.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5574–5584, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *PMLR*, pages 2649–2658. PMLR, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR 2014)*, 2014.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

Brenden M. Lake, Ruslan R. Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. doi: 10.1126/science.aab3050. Dataset URL `https://github.com/brendenlake/omniglot`.

Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pages 473–480, 2007. doi: 10.1145/1273496.1273556. Dataset URL `https://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/MnistVariations`.

Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems 2 (NIPS 1989)*, pages 396–404, 1990. Dataset URL `https://web.stanford.edu/~hastie/ElemStatLearn/data.html`.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. Dataset URL `http://yann.lecun.com/exdb/mnist/`.

James R. Lloyd and Zoubin Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 829–837, 2015.

Gaëlle Loosli, Stéphane Canu, and Léon Bottou. Training invariant support vector machines using selective sampling. In Léon Bottou, Olivier Chapelle, Dennis DeCoste, and Jason Weston, editors, *Large Scale Kernel Machines*, pages 301–320. MIT Press, Cambridge, MA, USA, 2007. Dataset URL `http://leon.bottou.org/projects/infimnist`.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. In *International Conference on Learning Representations Workshop (ICLRW 2016)*, 2016.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing sprites dataset, 2017. URL `https://github.com/deepmind/dsprites-dataset/`.

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *Proceedings of the Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2009)*, pages 296–301. IEEE, 2009. doi: 10.1109/AVSS.2009.58. Dataset URL `https://faces.dmi.unibas.ch/bfm/index.php?nav=1-1-1&id=scans`.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 2234–2242, 2016.

David W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization.* John Wiley & Sons, Inc., New York, NY, USA, 1992.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, pages 560–569. AUAI Press, 2018.

Dougal J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alexander J. Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations (ICLR 2017)*, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 2818–2826. IEEE, 2016. doi: 10.1109/CVPR.2016.308.

Loo-Nin Teow and Kia-Fock Loe. Robust vision-based features and classification schemes for off-line handwritten digit recognition. *Pattern Recognition*, 35(11):2355–2364, 2002. doi: 10.1016/S0031-3203(01)00228-X.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR 2016)*, 2016.

Tijmen Tieleman. affNIST, aug 2013. URL `https://www.cs.toronto.edu/~tijmen/affNIST/`.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in Python. *PeerJ*, 2:e453, 2014. doi: 10.7717/peerj.453.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017. arXiv:1708.07747. Dataset URL `https://github.com/zalandoresearch/fashion-mnist`.

Chhavi Yadav and Léon Bottou. Cold case: The lost MNIST digits, 2019. arXiv:1905.10498. Dataset URL `https://github.com/facebookresearch/qmnist`.