

APPENDIX

A PROOF OF THEOREM 1

Theorem 1 Assume the augmented data of each data point is obtained by the model defined according to equation (1)-(3) with parameters $\theta = (\mathbf{f}, \mathbf{T}, \lambda)$, and the following holds:

- (i) $\{\mathbf{a} \in \mathcal{A} | \varphi_\varepsilon(\mathbf{a}) = 0\}$ has measure zero, where φ_ε is the characteristic function of p_ε .
- (ii) The sufficient statistics $T_{i,j}$ are differentiable almost everywhere, and elements in \mathbf{T}_i are linearly independent on any subset of \mathcal{Z} with non-zero measure, $\forall i \in [n], j \in [k]$.
- (iii) λ is differentiable, and there exists a data point $\mathbf{x}^{(0)}$ such that the Jacobians of λ on it $J_\lambda(\mathbf{x}^{(0)})$ is row full rank.

then if there exists another model defined by Eqs. (1)-(3) with parameter $\tilde{\theta} = (\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\lambda})$. we have:

$$\mathbf{T}(\mathbf{f}^{-1}(\mathbf{a})) = \mathbf{A}\tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{a})) + \mathbf{c}, \forall \mathbf{a} \in \mathcal{A} \quad (8)$$

where \mathbf{A} and \mathbf{c} are constants, and \mathbf{A} is invertible.

proof. The proof consists of three steps, in which the first step and the third step are quoted from (Khemakhem et al., 2020).

Step I. Suppose there exists two sets of parameters $\theta = (\mathbf{f}, \mathbf{T}, \lambda)$ and $\tilde{\theta} = (\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\lambda})$ such that $p_{\mathbf{f}, \mathbf{T}, \lambda}(\mathbf{a}|\mathbf{x}) = p_{\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\lambda}}(\mathbf{a}|\mathbf{x}), \forall (\mathbf{a}, \mathbf{x}) \in (\mathcal{A}, \mathcal{X})$. Then:

$$\int_{\mathcal{Z}} p_{\mathbf{f}}(\mathbf{a}|\mathbf{z}) p_{\mathbf{T}, \lambda}(\mathbf{z}|\mathbf{x}) d\mathbf{z} = \int_{\mathcal{Z}} p_{\tilde{\mathbf{f}}}(\mathbf{a}|\mathbf{z}) p_{\tilde{\mathbf{T}}, \tilde{\lambda}}(\mathbf{z}|\mathbf{x}) d\mathbf{z} \quad (9)$$

Substitute the expression of $p_{\mathbf{f}}(\mathbf{a}|\mathbf{z})$:

$$\int_{\mathcal{Z}} p_\varepsilon(\mathbf{a} - \mathbf{f}(\mathbf{z})) p_{\mathbf{T}, \lambda}(\mathbf{z}|\mathbf{x}) d\mathbf{z} = \int_{\mathcal{Z}} p_\varepsilon(\mathbf{a} - \tilde{\mathbf{f}}(\mathbf{z})) p_{\tilde{\mathbf{T}}, \tilde{\lambda}}(\mathbf{z}|\mathbf{x}) d\mathbf{z} \quad (10)$$

Transform \mathbf{z} into \mathcal{A} , denoted as $\bar{\mathbf{a}}$:

$$\int_{\mathcal{A}} p_\varepsilon(\mathbf{a} - \bar{\mathbf{a}}) p_{\mathbf{T}, \lambda}(\mathbf{f}^{-1}(\bar{\mathbf{a}})|\mathbf{x}) \text{vol} J_{\mathbf{f}^{-1}}(\bar{\mathbf{a}}) d\bar{\mathbf{a}} = \int_{\mathcal{A}} p_\varepsilon(\mathbf{a} - \bar{\mathbf{a}}) p_{\tilde{\mathbf{T}}, \tilde{\lambda}}(\tilde{\mathbf{f}}^{-1}(\bar{\mathbf{a}})|\mathbf{x}) J_{\tilde{\mathbf{f}}^{-1}}(\bar{\mathbf{a}}) d\bar{\mathbf{a}} \quad (11)$$

This is a convolution on \mathbf{a} , and the kernel is $p_\varepsilon(\mathbf{a})$. Use Fourier transformation, and the transformed \mathbf{a} is denoted as w :

$$F[p_{\mathbf{T}, \lambda}(\mathbf{f}^{-1}(\mathbf{a})|\mathbf{x}) \text{vol} J_{\mathbf{f}^{-1}}(\mathbf{a})] \varphi_\varepsilon(w) = F[p_{\tilde{\mathbf{T}}, \tilde{\lambda}}(\tilde{\mathbf{f}}^{-1}(\mathbf{a})|\mathbf{x}) J_{\tilde{\mathbf{f}}^{-1}}(\mathbf{a})] \varphi_\varepsilon(w) \quad (12)$$

According to condition (i), $\varphi_\varepsilon(w) \neq 0$ almost every where. Hence we have:

$$F[p_{\mathbf{T}, \lambda}(\mathbf{f}^{-1}(\mathbf{a})|\mathbf{x}) \text{vol} J_{\mathbf{f}^{-1}}(\mathbf{a})] = F[p_{\tilde{\mathbf{T}}, \tilde{\lambda}}(\tilde{\mathbf{f}}^{-1}(\mathbf{a})|\mathbf{x}) J_{\tilde{\mathbf{f}}^{-1}}(\mathbf{a})] \quad (13)$$

Use the inverse Fourier transformation, we finally obtain:

$$p_{\mathbf{T}, \lambda}(\mathbf{f}^{-1}(\mathbf{a})|\mathbf{x}) \text{vol} J_{\mathbf{f}^{-1}}(\mathbf{a}) = p_{\tilde{\mathbf{T}}, \tilde{\lambda}}(\tilde{\mathbf{f}}^{-1}(\mathbf{a})|\mathbf{x}) J_{\tilde{\mathbf{f}}^{-1}}(\mathbf{a}) \quad (14)$$

This result shows that noise fulfilled condition (i) makes no difference theoretically.

Step II. Substitute the expression of $p_{\mathbf{T}, \lambda}(\mathbf{f}^{-1}(\mathbf{a})|\mathbf{x})$ in Eq. (14), and take the log on the both sides:

$$\begin{aligned} & \log \text{vol} J_{\mathbf{f}^{-1}}(\mathbf{a}) + \sum_{i=1}^n (\log Q_i(f_i^{-1}(\mathbf{a})) - \log Z_i(\lambda_i(\mathbf{x})) + \mathbf{T}_i(f_i^{-1}(\mathbf{a}))^\top \lambda_i(\mathbf{x})) \\ &= \log \text{vol} J_{\tilde{\mathbf{f}}^{-1}}(\mathbf{a}) + \sum_{i=1}^n (\log \tilde{Q}_i(\tilde{f}_i^{-1}(\mathbf{a})) - \log \tilde{Z}_i(\tilde{\lambda}_i(\mathbf{x})) + \tilde{\mathbf{T}}_i(\tilde{f}_i^{-1}(\mathbf{a}))^\top \tilde{\lambda}_i(\mathbf{x})) \end{aligned} \quad (15)$$

Take the first derivative with respect to \mathbf{x} :

$$J_\lambda(\mathbf{x})^\top \mathbf{T}(\mathbf{f}^{-1}(\mathbf{a})) - \sum_{i=1}^n J_{\lambda_i}(\mathbf{x})^\top \nabla_{\lambda_i} \log Z_i(\lambda_i) = J_{\tilde{\lambda}}(\mathbf{x})^\top \tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{a})) - \sum_{i=1}^n J_{\tilde{\lambda}_i}(\mathbf{x})^\top \nabla_{\tilde{\lambda}_i} \log \tilde{Z}_i(\tilde{\lambda}_i) \quad (16)$$

Table 1: Ranges of three transformations.

Transformation	Range
translation	$[-0.1, 0.1] \times [-0.1, 0.1]$
rotation	$[0, 5^\circ]$
scaling	$[0.8, 1.25]$

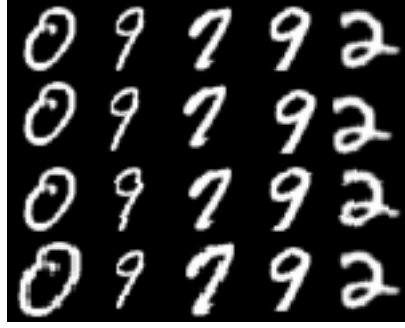


Figure 3: **Five samples and their augmented images by the three transformations.** The first row shows the original images. The next three rows show the images augmented by horizontal and vertical translation, rotation and scaling, respectively.

According to condition (iii), let $\mathbf{x} = \mathbf{x}^{(0)}$, then $J_{\lambda}(\mathbf{x}^{(0)})^{\top}$ is row full rank, and hence has left pseudo inverse matrix: $(J_{\lambda}(\mathbf{x}^{(0)})J_{\lambda}(\mathbf{x}^{(0)})^{\top})^{-1}J_{\lambda}(\mathbf{x}^{(0)})$. Multiply the both sides by the left pseudo inverse matrix, we have:

$$\mathbf{T}(\mathbf{f}^{-1}(\mathbf{a})) = \mathbf{A}\tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{a})) + \mathbf{c} \quad (17)$$

where:

$$\begin{aligned} \mathbf{A} &= (J_{\lambda}(\mathbf{x}^{(0)})J_{\lambda}(\mathbf{x}^{(0)})^{\top})^{-1}J_{\lambda}(\mathbf{x}^{(0)})J_{\tilde{\lambda}}(\mathbf{x}^{(0)})^{\top} \\ \mathbf{c} &= (J_{\lambda}(\mathbf{x}^{(0)})J_{\lambda}(\mathbf{x}^{(0)})^{\top})^{-1}J_{\lambda}(\mathbf{x}^{(0)}) \sum_{i=1}^n (J_{\lambda_i}(\mathbf{x}^{(0)})^{\top} \nabla_{\lambda_i} \log Z_i(\lambda_i) - J_{\tilde{\lambda}_i}(\mathbf{x}^{(0)})^{\top} \nabla_{\tilde{\lambda}_i} \log \tilde{Z}_i(\tilde{\lambda}_i)) \end{aligned} \quad (18)$$

Note that $p_{\mathbf{T}, \lambda}(\mathbf{z}|\mathbf{x})$ has support cross the entire latent space \mathbb{R}^n , and hence $p_{\mathbf{f}, \mathbf{T}, \lambda}(\mathbf{a}|\mathbf{x}^0)$ has support cross \mathcal{A} . As a result, Eq. 17 holds for any $\mathbf{a} \in \mathcal{A}$. This means that \mathbf{A} and \mathbf{c} are independent of \mathbf{x} .

Step III. In this step, we prove the invertibility of \mathbf{A} . Let $\mathbf{z} = \mathbf{f}^{-1}(\mathbf{a})$, then Eq. 17 can be written as: $\mathbf{T}(\mathbf{z}) = \mathbf{A}\tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1} \circ \mathbf{f}(\mathbf{z})) + \mathbf{c}$. Take the first derivative with respect to \mathbf{z} , then: $J_{\mathbf{T}}(\mathbf{z}) = \mathbf{A}J_{\mathbf{T} \circ \mathbf{f}^{-1} \circ \mathbf{f}}(\mathbf{z})$. According to the Lemma by (Khemakhem et al., 2020), by condition (ii), there exists a k distinct latent samples $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(k)}\}$ such that is $(J_{\mathbf{T}}(\mathbf{z}^{(1)}), \dots, J_{\mathbf{T}}(\mathbf{z}^{(k)}))$ is invertible. Note that $(J_{\mathbf{T}}(\mathbf{z}^{(1)}), \dots, J_{\mathbf{T}}(\mathbf{z}^{(k)})) = \mathbf{A}(J_{\mathbf{T} \circ \mathbf{f}^{-1} \circ \mathbf{f}}(\mathbf{z}^{(1)}), \dots, J_{\mathbf{T} \circ \mathbf{f}^{-1} \circ \mathbf{f}}(\mathbf{z}^{(k)}))$, hence \mathbf{A} is invertible.

B AUGMENTATIONS ON EMNIST

The augmentations used on EMNIST in this work includes three affine transformations: horizontal and vertical translation, rotation and scaling. Their parameters are randomly chosen from a uniform distribution on certain ranges for each data point. The ranges are shown in Table B. Five images augmented by these transformations are shown in Fig. 3. Note that in experiments, we use the three transformations and their combinations, totally 7 types, for augmentations.

C FIGURES FROM THE EMNIST EXPERIMENTS

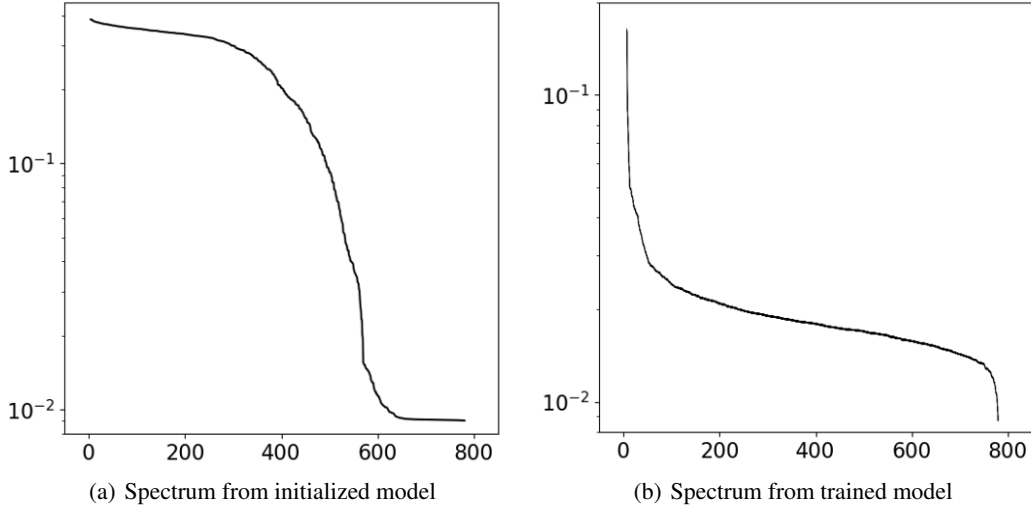


Figure 4: Spectrum of sorted standard deviations from model initialized by identity transformation and the trained model. X-axis is sorted latent dimensions; y-axis is standard deviations in log scale.

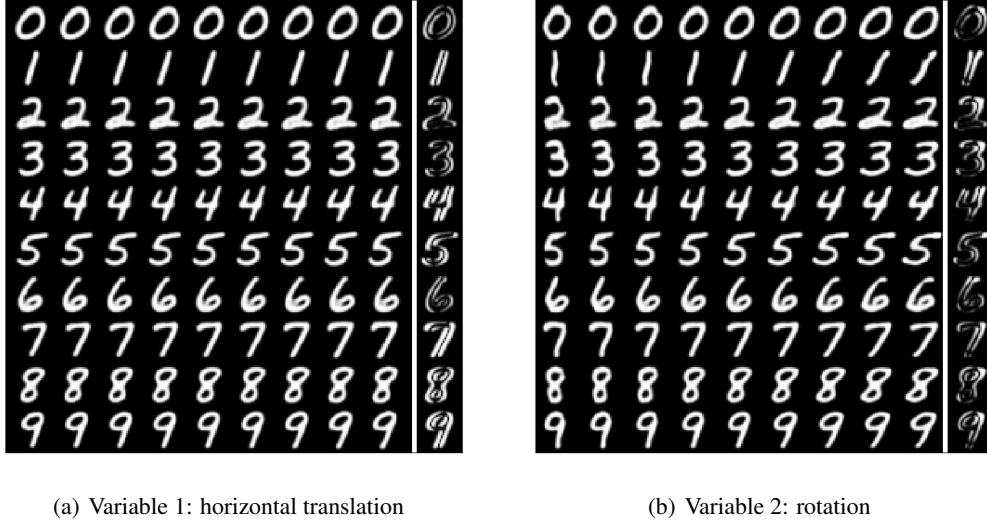


Figure 5: Most significant latent variables 1-2. Each row is generated by three steps: i) Samples in EMNIST Digits testing set with corresponding label are encoded into estimated latent space by the trained model. ii) Each estimated latent variable is averaged. iii) Manipulate the chosen latent variable on $[-2, 2]$ and then decode the latent variables by the reverse trained model. The rightmost column show the areas affected by the chosen latent variable, computing by the absolute pixel difference on $[-1, 1]$.



(a) Variable 3: height of the top half



(b) Variable 4: width of the bottom half by the bottom right corner

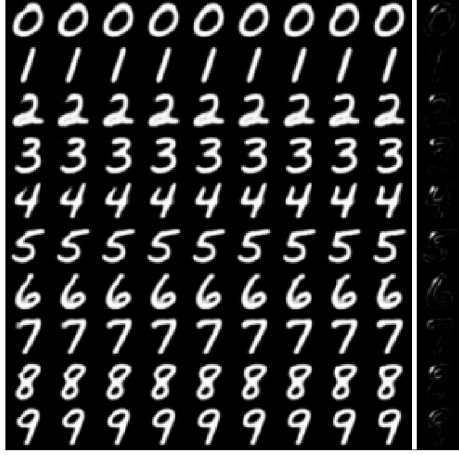


(c) Variable 5: height of the bottom half



(d) Variable 6: black space of the lower half

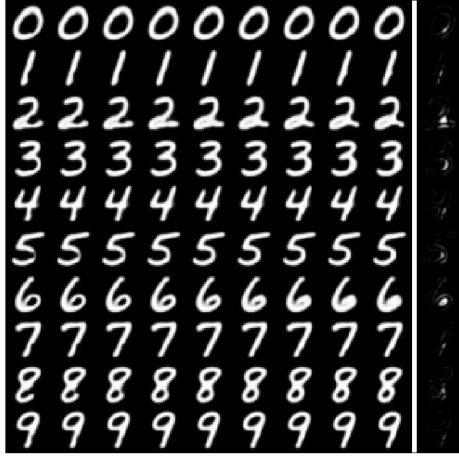
Figure 6: Most significant latent variables 3-6. Each row is generated by three steps: i) Samples in EMNIST Digits testing set with corresponding label are encoded into estimated latent space by the trained model. ii) Each estimated latent variable is averaged. iii) Manipulate the chosen latent variable on $[-2, 2]$ and then decode the latent variables by the reverse trained model. The rightmost column show the areas affected by the chosen latent variable, computing by the absolute pixel difference on $[-1, 1]$.



(a) Variable 7: width of the top half by the top left corner



(b) Variable 8: width of the top half by the lower top left corner

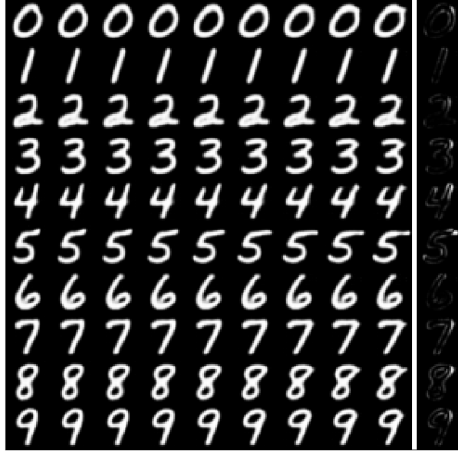


(c) Variable 9: openness of the bottom half



(d) Variable 10: openness of the top half

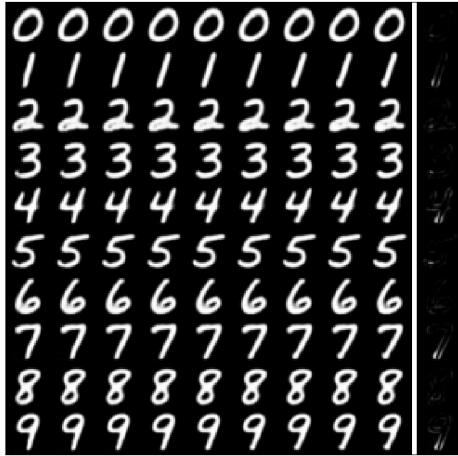
Figure 7: Most significant latent variables 7-10. Each row is generated by three steps: i) Samples in EMNIST Digits testing set with corresponding label are encoded into estimated latent space by the trained model. ii) Each estimated latent variable is averaged. iii) Manipulate the chosen latent variable on $[-2, 2]$ and then decode the latent variables by the reverse trained model. The rightmost column show the areas affected by the chosen latent variable, computing by the absolute pixel difference on $[-1, 1]$.



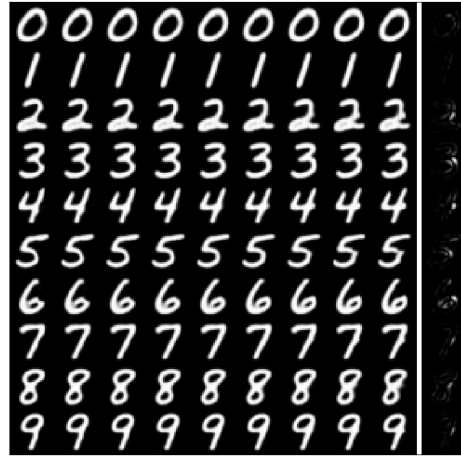
(a) Variable 11: top right corner



(b) Variable 12: lower top left corner



(c) Variable 13: bend of vertical bar in 1, 4, 7, 9



(d) Variable 14: extension of right corner

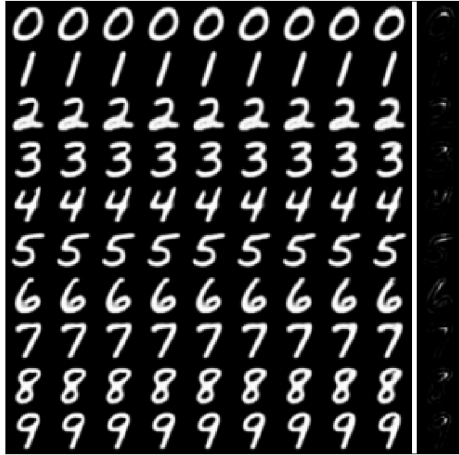
Figure 8: Most significant latent variables 11-14. Each row is generated by three steps: i) Samples in EMNIST Digits testing set with corresponding label are encoded into estimated latent space by the trained model. ii) Each estimated latent variable is averaged. iii) Manipulate the chosen latent variable on $[-2, 2]$ and then decode the latent variables by the reverse trained model. The rightmost column show the areas affected by the chosen latent variable, computing by the absolute pixel difference on $[-1, 1]$.



(a) Variable 15: extension of higher right corner



(b) Variable 16: extension of lower right corner

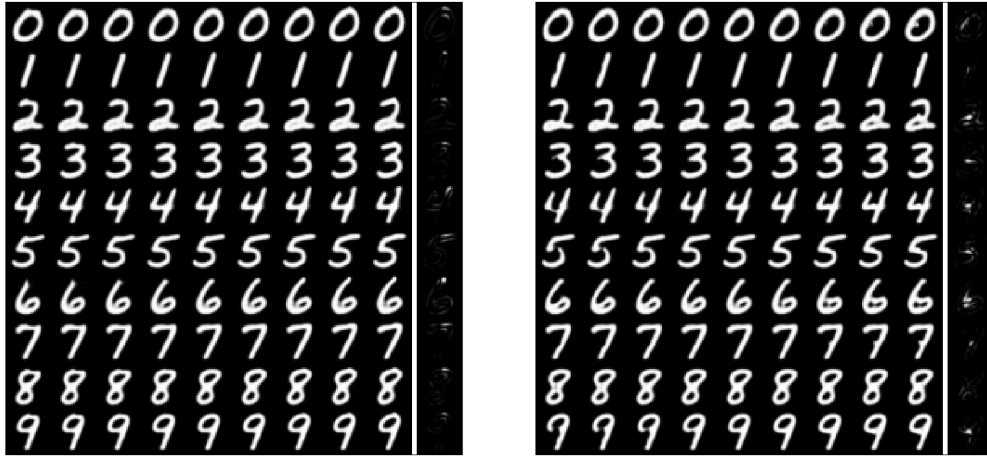


(c) Variable 17: size of top capity by lower top right corner



(d) Variable 18: size of top capity by lower top left corner

Figure 9: Most significant latent variables 15-18. Each row is generated by three steps: i) Samples in EMNIST Digits testing set with corresponding label are encoded into estimated latent space by the trained model. ii) Each estimated latent variable is averaged. iii) Manipulate the chosen latent variable on $[-2, 2]$ and then decode the latent variables by the reverse trained model. The rightmost column show the areas affected by the chosen latent variable, computing by the absolute pixel difference on $[-1, 1]$.



(a) Variable 19: bend of the top horizontal bar

(b) Variable 20: existence of the crossbar

Figure 10: Most significant latent variables 19-20. Each row is generated by three steps: i) Samples in EMNIST Digits testing set with corresponding label are encoded into estimated latent space by the trained model. ii) Each estimated latent variable is averaged. iii) Manipulate the chosen latent variable on $[-2, 2]$ and then decode the latent variables by the reverse trained model. The rightmost column show the areas affected by the chosen latent variable, computing by the absolute pixel difference on $[-1, 1]$.