

Quantitative Understanding of VAE as a Non-linearly Scaled Isometric Embedding

Akira Nakagawa¹, Keizo Kato¹, Taiji Suzuki^{2, 3}

¹ Fujitsu Limited

² The University of Tokyo

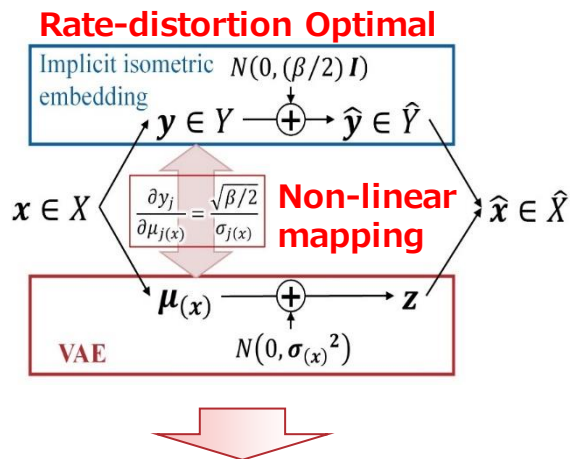
³ RIKEN

Overview

Clear interpretation of VAE using both differential-geometric and information-theoretic frameworks.

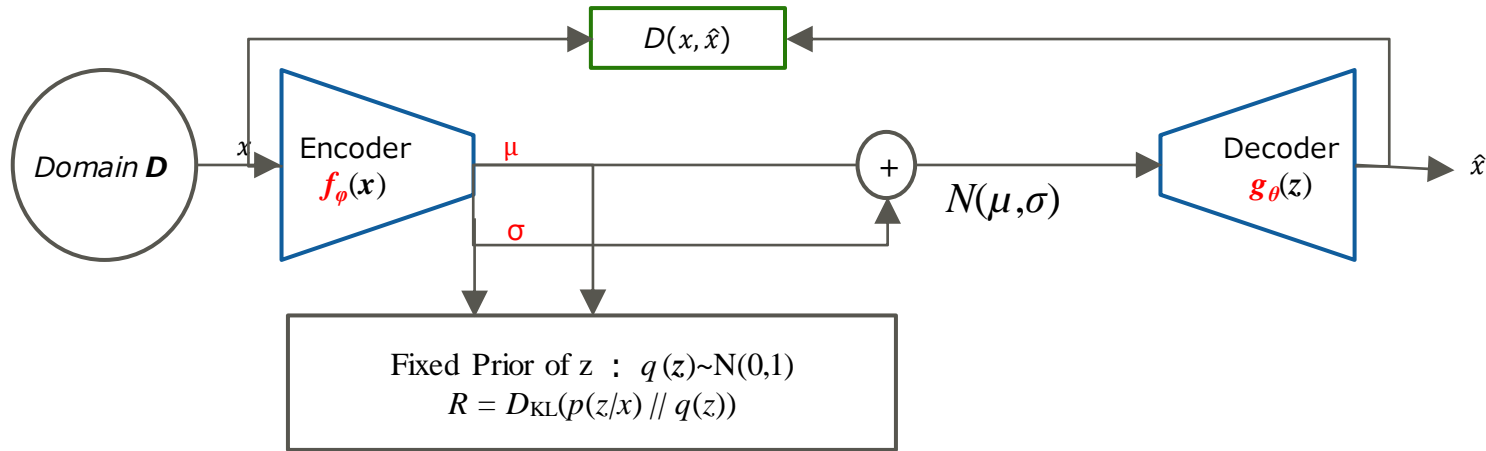
- Introduce the isometric embedding from the metric-defined input space to the Euclidean implicit latent space.
- VAE achieves the Rate-distortion optimal in the implicit isometric latent space.
- Isometricity allows quantitative data analysis such as PDF estimation, anomaly detection, and PCA-like analysis.

Mapping VAE to implicit isometric embedding.



Quantitative Data Analysis

β -VAE



$$\theta, \varphi = \arg \min (E_{x, \sigma} [D(x, \hat{x}) + \beta \cdot D_{\text{KL}}(p(z|x) || q(z))])$$

The objective is a summation of the reconstruction loss and KL divergence.

- Reconstruction Loss between input and reconstruction data, formulated as:
 $-\log(p_\theta(x|z)) = -\log(p_\theta(x|\hat{x})) = D(x, \hat{x}) \approx (x - \hat{x})^T G_x (x - \hat{x})$ where G_x is a metric tensor
- KL divergence between prior and posterior

Introduction of Implicit Isometric embedding

Derivation and Features

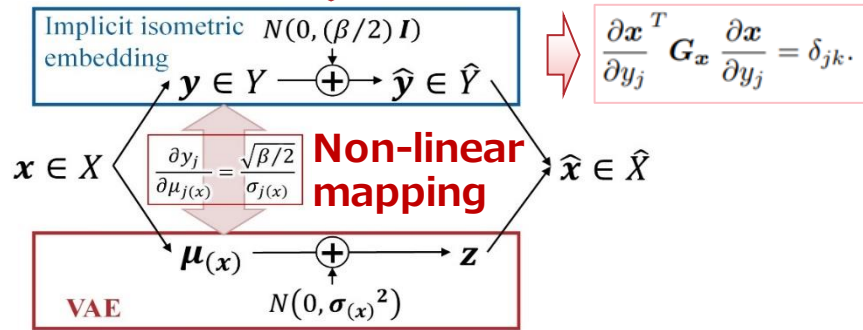
- Map μ to y such that $dy_j/d\mu_j = \sqrt{\beta/2}/\sigma_j$

Features of y

- Isometric Embedding of input data
- Constant posterior variance $\beta/2$, resulting Rate-Distortion optimal

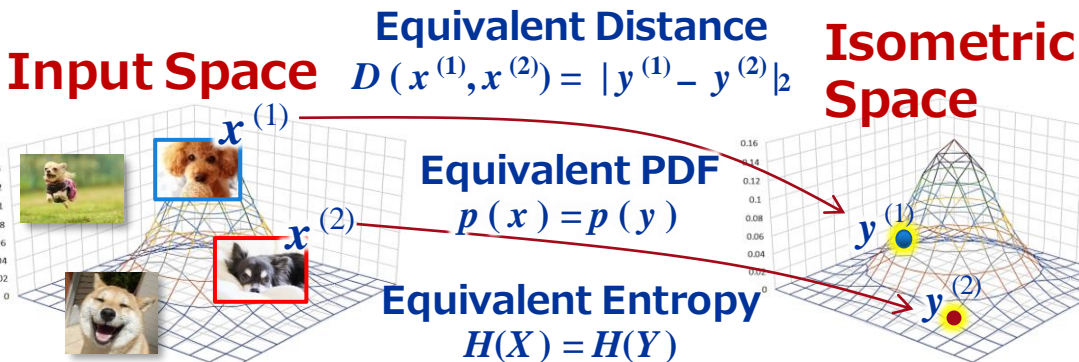
Constant Posterior Variance (RD optimal)

Isometric Property



Preferable properties

- Equivalent Distance, PDF, and Entropy in both spaces.



Proof Outline

- Approximation of the objective where \mathbf{x}_{μ_j} denotes $\partial \mathbf{x} / \partial \mu_j(\mathbf{x})$.

$$L_{\mathbf{x}} \simeq \sum_{j=1}^n \sigma_{j(\mathbf{x})}^2 {}^t \mathbf{x}_{\mu_j} \mathbf{G}_x \mathbf{x}_{\mu_j} - \beta \log \left(p(\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}(\mathbf{x})} \right) \right| \prod_{j=1}^n \sigma_{j(\mathbf{x})} \right) - \frac{n\beta \log 2\pi e}{2}$$

- Derivation of $L_{\mathbf{x}}$ by \mathbf{x}_{μ_j} where $\tilde{\mathbf{x}}_{\mu_j}$ denotes the column vector in the cofactor matrix of Jacobian $\frac{\partial \mathbf{x}}{\partial \boldsymbol{\mu}(\mathbf{x})}$.

$$\frac{\partial L_{\mathbf{x}}}{\partial \mathbf{x}_{\mu_j}} = 2\sigma_{j(\mathbf{x})}^2 \mathbf{G}_x \mathbf{x}_{\mu_j} - \frac{\beta}{\det(\partial \mathbf{x} / \partial \boldsymbol{\mu}(\mathbf{x}))} \tilde{\mathbf{x}}_{\mu_j} = 0$$

⇒ $(2\sigma_{j(\mathbf{x})}^2 / \beta) {}^t \mathbf{x}_{\mu_k} \mathbf{G}_x \mathbf{x}_{\mu_j} = \delta_{jk}$ (Orthogonal in the metric space of G_x)

- Map $\boldsymbol{\mu}$ to \mathbf{y} such that $dy_j/d\mu_j = \sqrt{\beta/2}/\sigma_j$.

⇒ ${}^t \mathbf{x}_{y_j} \mathbf{G}_x \mathbf{x}_{y_k} = \delta_{jk}$ (Isometric in the metric space of G_x)

Quantitative Properties of VAE

■ Isometricity

- Show Unit norm of implicit isometric component

$${}^t\mathbf{x}_{y_j} \mathbf{G}_x \mathbf{x}_{y_j} \simeq (2/\beta) (\sigma_{j(\mathbf{x})})^2 {}^t\mathbf{x}_{\mu_j} \mathbf{G}_x \mathbf{x}_{\mu_j} \simeq 1$$

■ Probability estimation

- Isometricity enables probability estimation of input data, where $\sigma_{j(\mathbf{x})}$ bridges the distributions between the input and prior.

$$p_{\mathbf{G}_x}(\mathbf{x}) \simeq p(\mathbf{y}) \propto p(\boldsymbol{\mu}(\mathbf{x})) \prod_{j=1}^m \sigma_{j(\mathbf{x})}$$

■ Importance of each latent component

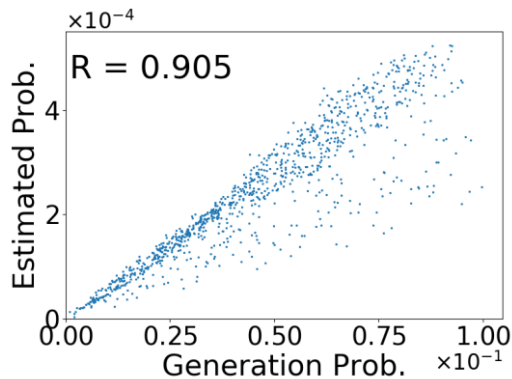
- PCA eigenvalue-like importance of each component can be evaluated

$$\text{Var}(y_j) \simeq (\beta/2) E_{\mathbf{x} \sim p(\mathbf{x})} [\sigma_{j(\mathbf{x})}^{-2}].$$

Quantitative data analysis

Implicit isometric embedding allows for quantitative data analysis

PDF estimation



Enable to estimates the input distribution $p(x)$

Anomaly detection

Dataset	Methods	F1
KDDCup	GMVAE*	0.9326
	DAGMM*	0.9500 (0.0052)
	RaDOGAGA(d)*	0.9624 (0.0038)
	RaDOGAGA(log(d))*	0.9638 (0.0042)
	vanilla VAE	0.9642 (0.0007)
Thyroid	GMVAE*	0.6353
	DAGMM*	0.4755 (0.0491)
	RaDOGAGA(d)*	0.6447 (0.0486)
	RaDOGAGA(log(d))*	0.6702 (0.0585)
	vanilla VAE	0.6596 (0.0436)
Arrythmia	GMVAE*	0.4308
	DAGMM*	0.5060 (0.0395)
	RaDOGAGA(d)*	0.5433 (0.0468)
	RaDOGAGA(log(d))*	0.5373 (0.0411)
	vanilla VAE	0.4985 (0.0412)
KDDCup-rev	DAGMM*	0.9779 (0.0018)
	RaDOGAGA(d)*	0.9797 (0.0015)
	RaDOGAGA(log(d))*	0.9865 (0.0009)
	vanilla VAE	0.9880 (0.0008)

Comparable to Previous SOTA

Importance of each VAE latent variable

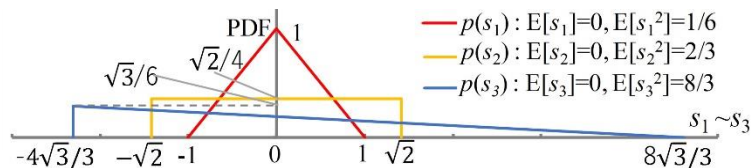


$E_{p(x)}[\sigma_j^{-2}]$ shows each dimensional importance

Probability estimation in Toy data

Configuration

- ① sample s_1 , s_2 , and s_3 from PDFs as below:



- ② Generate \mathbf{x} from s_1 , s_2 , and s_3 as:

$$\mathbf{x} = \sum_{i=1}^3 s_i \mathbf{v}_i$$

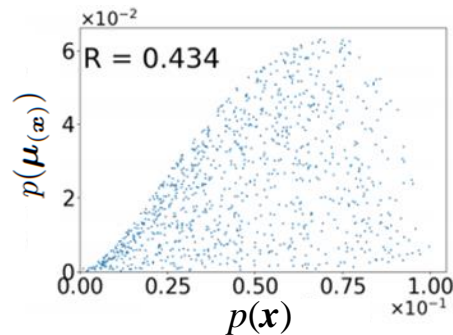
where \mathbf{v}_1 , \mathbf{v}_2 , and $\mathbf{v}_3 \in \mathbb{R}^{16}$ are uncorrelated vectors with unit norm. $p(\mathbf{x})$ is set to $p(s_1) p(s_2) p(s_3)$.

- ③ Train VAE for data \mathbf{x} . Then $p(\mathbf{x})$ and its estimations are plotted.

Probability estimation

PDF of Prior $p(\boldsymbol{\mu}(\mathbf{x}))$

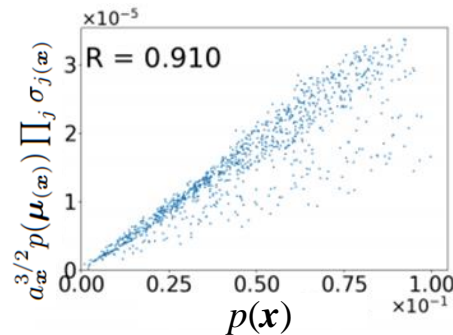
Poor estimation with low correlation $R=0.434$



PDF in isometric space

$$\frac{a_{\mathbf{x}}^{3/2} p(\boldsymbol{\mu}(\mathbf{x})) \prod_j \sigma_j(\mathbf{x})}{}$$

Good estimation with high correlation $R=0.910$



Anomaly Detection Task

By evaluating the data probabilities in the isometric space, even vanilla VAE is comparable to the previous SOTA (RaDOGAGA).

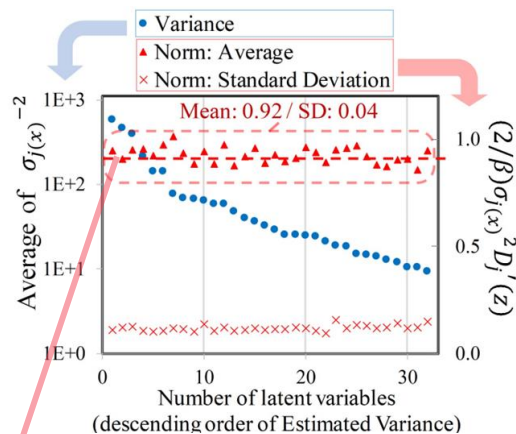
Table 6. Average and standard deviations (in brackets) of Precision, Recall and F1

Dataset	Methods	Precision	Recall	F1
KDDCup	GMVAE [†]	0.952	0.9141	0.9326
	DAGMM [†]	0.9427 (0.0052)	0.9575 (0.0053)	0.9500 (0.0052)
	RaDOGAGA(d) [†]	0.9550 (0.0037)	0.9700 (0.0038)	0.9624 (0.0038)
	RaDOGAGA(log(d)) [†]	0.9563 (0.0042)	0.9714 (0.0042)	0.9638 (0.0042)
	VAE	0.9568(0.0007)	0.9718 (0.0007)	0.9642(0.0007)
Thyroid	GMVAE [†]	0.7105	0.5745	0.6353
	DAGMM [†]	0.4656 (0.0481)	0.4859 (0.0502)	0.4755 (0.0491)
	RaDOGAGA(d) [†]	0.6313 (0.0476)	0.6587 (0.0496)	0.6447 (0.0486)
	RaDOGAGA(log(d)) [†]	0.6562 (0.0572)	0.6848 (0.0597)	0.6702 (0.0585)
	VAE	0.6458 (0.04270)	0.6739 (0.04455)	0.6596 (0.0436)
Arrythmia	GMVAE [†]	0.4375	0.4242	0.4308
	DAGMM [†]	0.4985 (0.0389)	0.5136 (0.0401)	0.5060 (0.0395)
	RaDOGAGA(d) [†]	0.5353 (0.0461)	0.5515 (0.0475)	0.5433 (0.0468)
	RaDOGAGA(log(d)) [†]	0.5294 (0.0405)	0.5455 (0.0418)	0.5373 (0.0411)
	VAE	0.4912(0.0406)	0.5061 (0.0419)	0.4985 (0.0413)
KDDCup-rev	DAGMM [†]	0.9778 (0.0018)	0.9779 (0.0017)	0.9779 (0.0018)
	RaDOGAGA(d) [†]	0.9768 (0.0033)	0.9827 (0.0012)	0.9797 (0.0015)
	RaDOGAGA(log(d)) [†]	0.9864 (0.0009)	0.9865 (0.0009)	0.9865 (0.0009)
	VAE	0.9880 (0.0008)	0.9881 (0.0008)	0.9880 (0.0008)

Scores are cited from Liao et al. (2018) (GMVAE) and Kato et al. (2020)(DAGMM, RaDOGAGA)

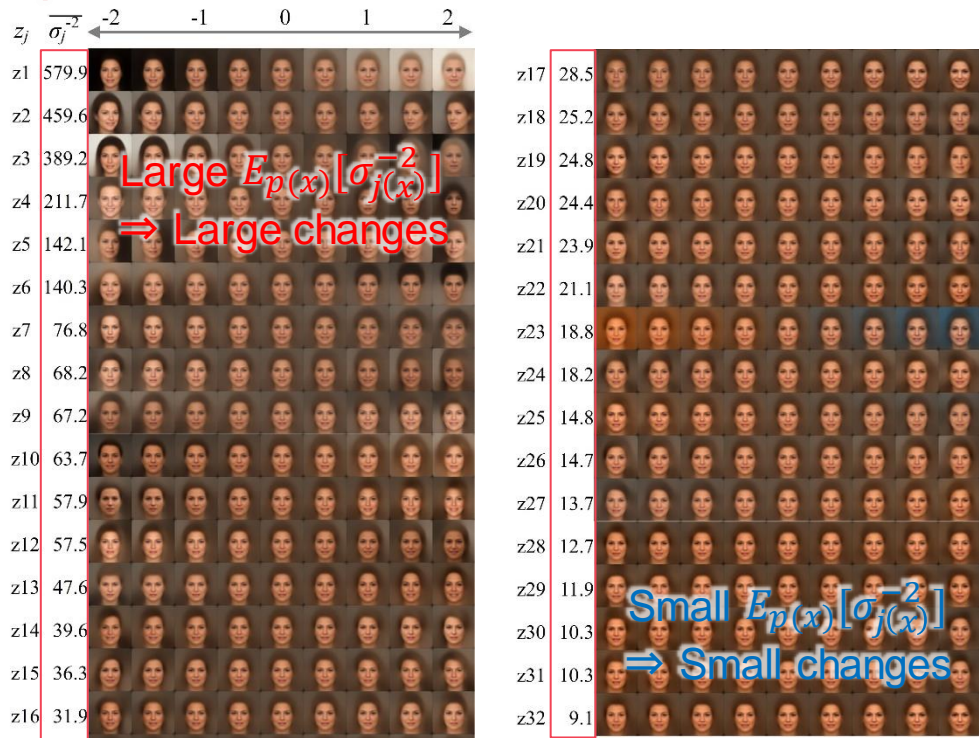
Celeb A analysis using VAE

Isometricity evaluation



${}^t x_{y_j} G_x x_{y_j} \simeq (2/\beta) (\sigma_{j(x)}^2 {}^t x_{\mu_j} G_x x_{\mu_j}) \simeq 1$
 almost holds, showing isometricity.

Importance estimation of latent variables

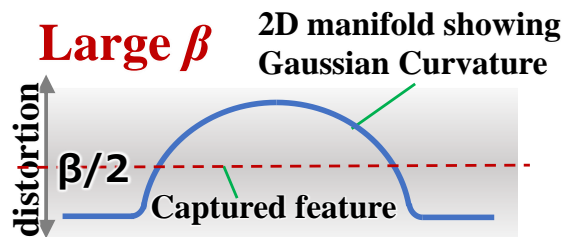


$E_p(x)[\sigma_j^{-2}]$ shows each dimensional importance.

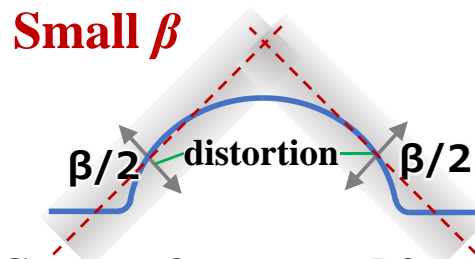
Explanation of VAE Behaviors when varying β

Our work gives an intuitive explanation why larger β can capture global features (Higgins et.al 2017)

- Assume 2D manifold in 3D space, where isometric embedding is governed by Gaussian curvature according to "Gauss's Theorema Egregium."
 - Large β : Capture global features allowing large distortion $\beta/2$
 - Small β : Capture only fragmented features allowing small distortion $\beta/2$
- ⇒ Similar behaviors will occur in higher dimensional manifold.



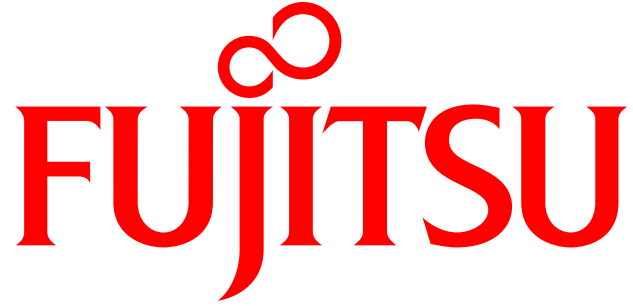
**Capture global features
allowing large distortion**



**Capture fragmented features
allowing small distortion.**

Conclusion

- Many prior works analyzed VAE, however, its quantitative behaviors have not been well clarified yet.
- We thoroughly revealed the quantitative behaviors of VAE in the differential-geometric and information-theoretic framework such that VAE achieves the Rate-distortion optimal embedding in the implicit isometric embedding.
- Our work will provide an important starting point for quantitative generative model to understand the real world.




shaping tomorrow with you

What is β -VAE

Correctly, what β -VAE really does is only to scale the variance of the pre-determined conditional distribution in the original VAE by a factor of β . In the case the pre-determined conditional distribution is Gaussian $\mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \sigma^2 \mathbf{I})$, the objective of β -VAE can be rewritten as a linearly scaled original VAE objective with a Gaussian $\mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \beta\sigma^2 \mathbf{I})$ where the variance is $\beta\sigma^2$ instead of σ^2 :

$$\begin{aligned} \underbrace{E_{q_\phi(\cdot)}[\log \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \sigma^2 \mathbf{I})] - \beta D_{\text{KL}}(\cdot)}_{\beta\text{-VAE objective with pre-determined conditional distribution in decoder}} &= E_{q_\phi(\cdot)} \left[-\frac{1}{2} \log 2\pi\sigma^2 - \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}{2\sigma^2} \right] - \beta D_{\text{KL}}(\cdot) \\ &= \beta \left(E_{q_\phi(\cdot)} \left[-\frac{1}{2} \log 2\pi\beta\sigma^2 - \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}{2\beta\sigma^2} \right] - D_{\text{KL}}(\cdot) \right) \\ &\quad + \frac{\beta}{2} \log 2\pi\beta\sigma^2 - \frac{1}{2} \log 2\pi\sigma^2 \\ &= \beta \left(\underbrace{E_{q_\phi(\cdot)}[\log \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \beta\sigma^2 \mathbf{I})] - D_{\text{KL}}(\cdot)}_{\text{Original VAE objective with scaled conditional distribution}} \right) + \text{const.} \end{aligned} \quad (81)$$



FUJITSU

shaping tomorrow with you