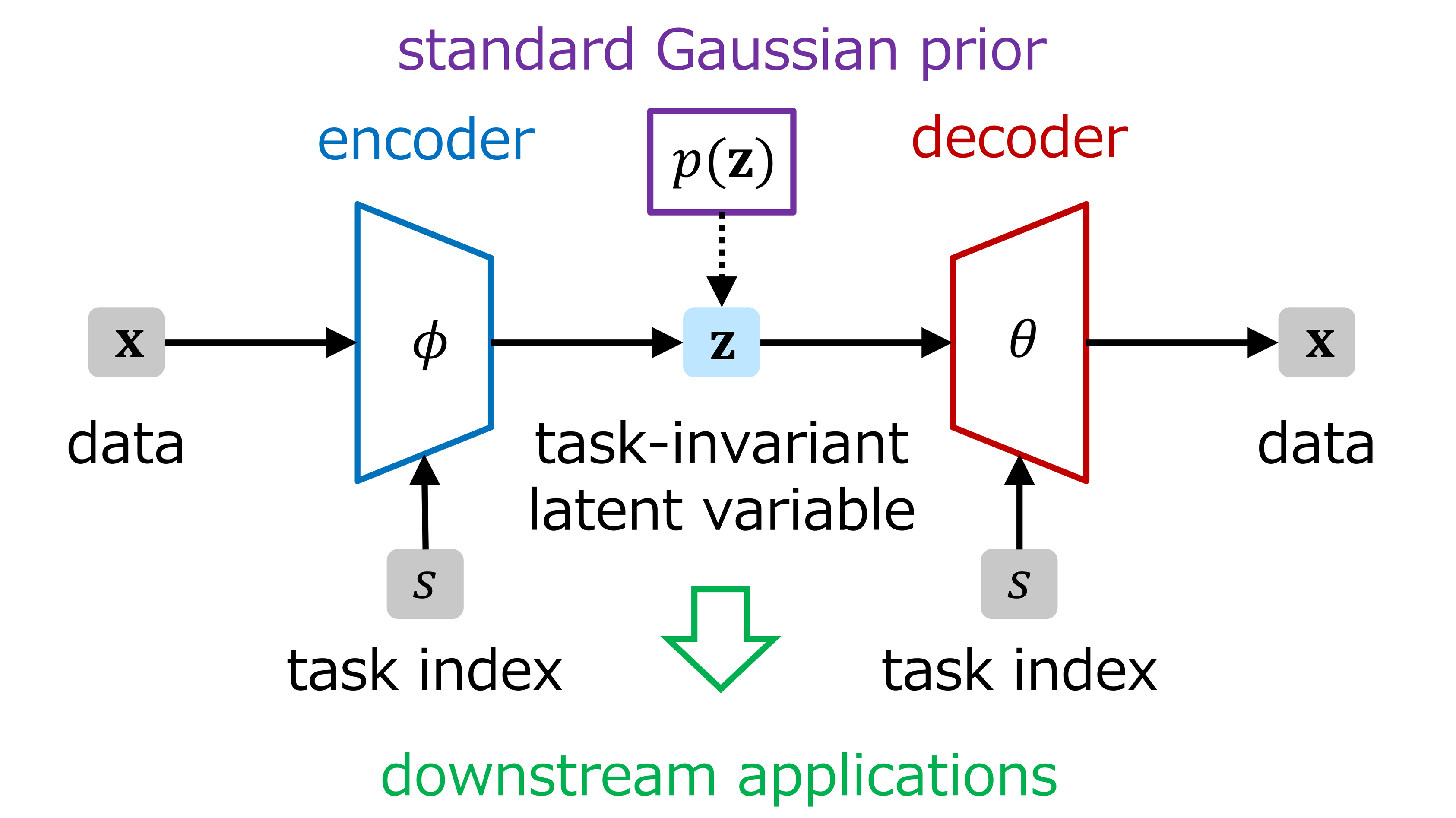
Learning Optimal Priors for Task-Invariant Representations in VAEs

Hiroshi Takahashi[1], Tomoharu Iwata[1], Atsutoshi Kumagai[1], Sekitoshi Kanai[1],
Masanori Yamada[1], Yuuki Yamanaka[1], Hisashi Kashima[2]
[1]NTT, [2]Kyoto University

1. Conditional Variational Autoencoder

- The variational autoencoder (VAE) is a powerful latent variable model for unsupervised representation learning, but it cannot perform well with insufficient data points.
- To solve this, the conditional VAE (CVAE) tries to obtain task-invariant latent variable from multiple tasks.



- The CVAE models a conditional probability of $\mathbf{x}$ given s as:

$$p_\theta(\mathbf{x}|s) = \int \underbrace{p_\theta(\mathbf{x}|\mathbf{z},s)}_{\text{decoder}}\underbrace{p(\mathbf{z})}_{\text{prior}}\mathrm{d}\mathbf{z} = \mathbb{E}_{\underbrace{q_\phi(\mathbf{z}|\mathbf{x},s)}_{\text{encoder}}}\left[\frac{p_\theta(\mathbf{x}|\mathbf{z},s)p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x},s)}\right]$$

- The CVAE is trained by maximizing the evidence lower bound (ELBO) as follows:

$$\mathcal{F}_{\mathrm{CVAE}}(\theta,\phi) = \mathbb{E}_{p_D(\mathbf{x},s)q_\phi(\mathbf{z}|\mathbf{x},s)}\left[\ln p_\theta(\mathbf{x}|\mathbf{z},s)\right] - \underbrace{\mathbb{E}_{p_D(\mathbf{x},s)}\left[D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},s)\|p(\mathbf{z}))\right]}_{=\mathcal{R}(\phi)}$$

data distribution

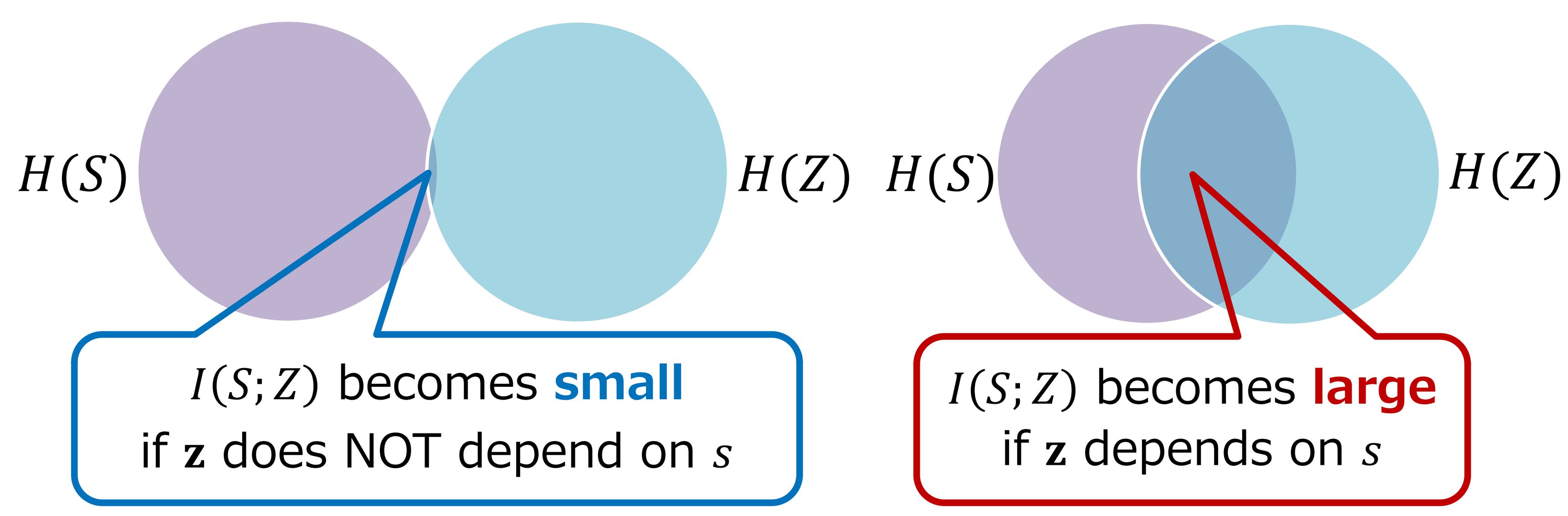
2. Problem and Contribution

- Although the CVAE can reduce the dependency of $\mathbf{z}$ on s to some extent, this dependency remains in many cases.
- The contribution of this study is as follows:

1. We investigate the cause of the task-dependency in the CVAE and reveal the **simple prior** is one of the causes.
2. We introduce the **optimal prior** to reduce the task-dependency.
3. We theoretically and experimentally show that our learned representation works well on multiple tasks.

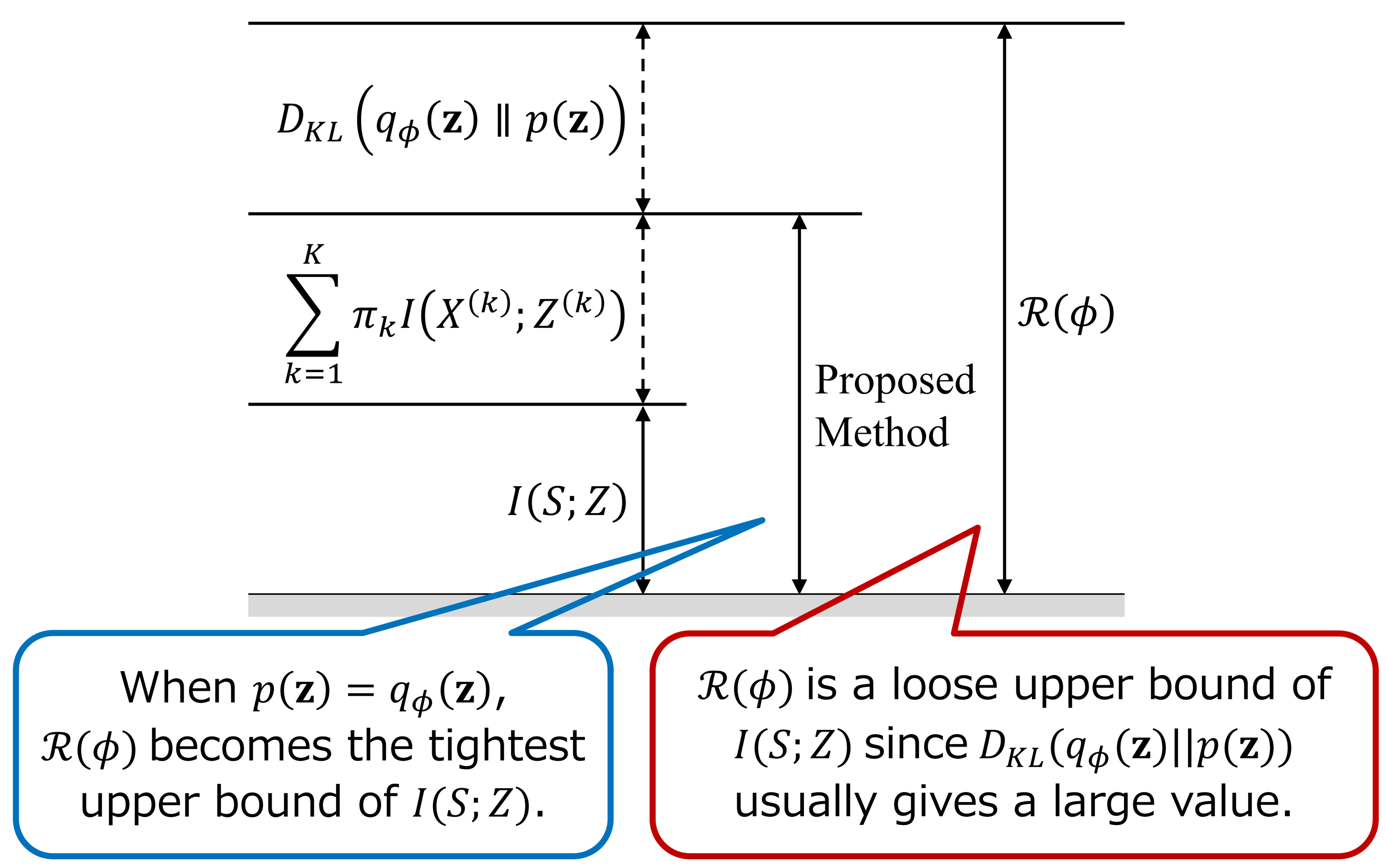
3. Mutual Information

- To investigate the cause of dependency of $\mathbf{z}$ on s, we introduce the mutual information $I(S;Z)$ that measures the mutual dependence between two random variables.



4a. Theorem 1

- The CVAE tries to minimize the mutual information $I(S;Z)$ by minimizing its **loose** upper bound $\mathcal{R}(\phi)$:



- That is, the simple prior $p(\mathbf{z})$ is **one causes of the task-dependency**, and $q_\phi(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{x},s)p_D(\mathbf{x},s)\mathrm{d}\mathbf{x}$ is the **optimal prior** to reduce it.

4b. Theorem 2

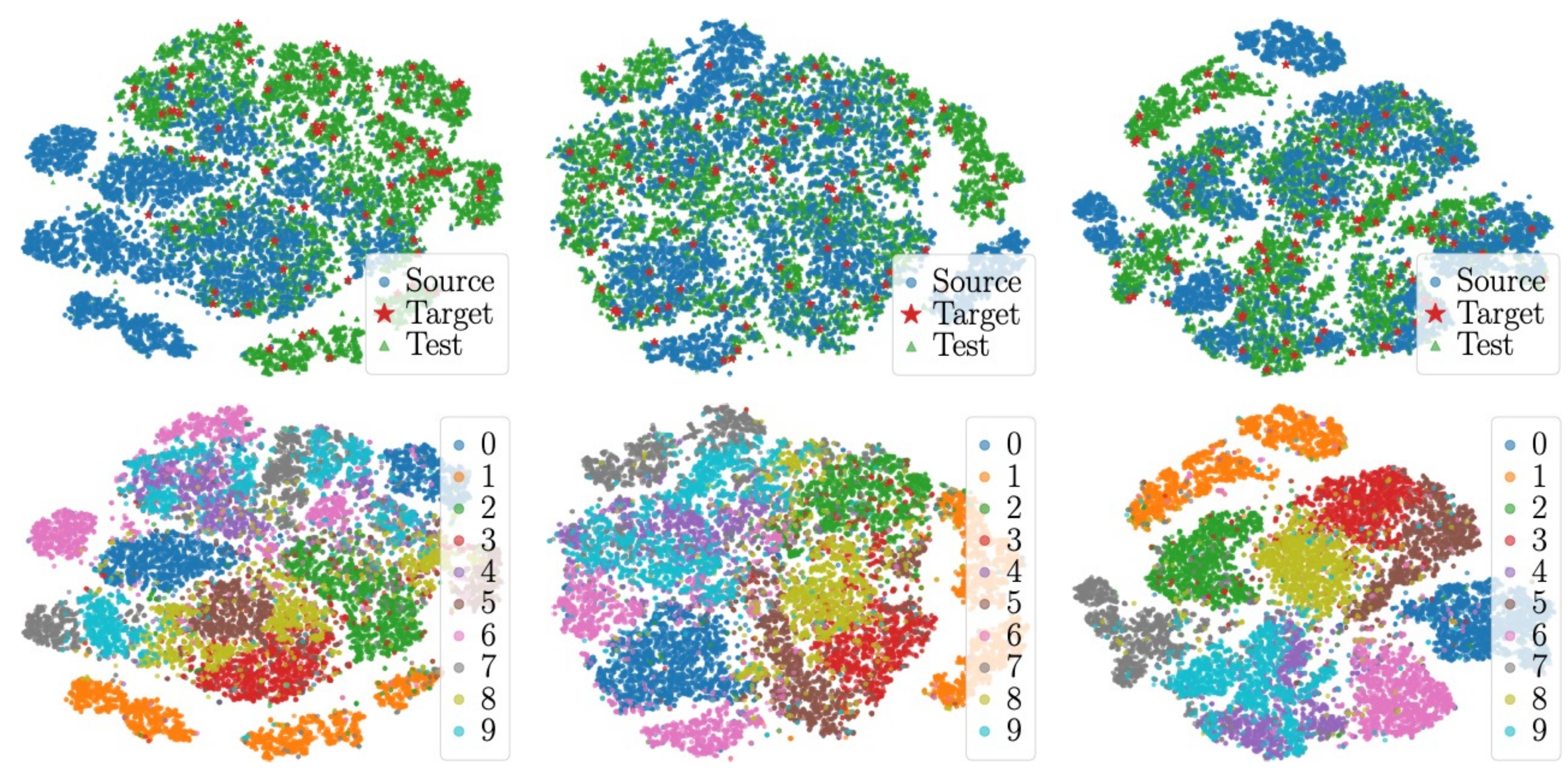
- The ELBO with this optimal prior $\mathcal{F}_{\mathrm{Proposd}}(\theta,\phi)$ is **always larger than** or equal to original ELBO $\mathcal{F}_{\mathrm{CVAE}}(\theta,\phi)$:

$$\mathcal{F}_{\mathrm{Proposed}}(\theta,\phi) = \mathcal{F}_{\mathrm{CVAE}}(\theta,\phi) + D_{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z})) \geq \mathcal{F}_{\mathrm{CVAE}}(\theta,\phi)$$

- That is, $\mathcal{F}_{\mathrm{Proposd}}(\theta,\phi)$ is also a **better lower bound of the log-likelihood** than $\mathcal{F}_{\mathrm{CVAE}}(\theta,\phi)$, which contributes to obtaining better representation.

5. Experiments

Visualization of latent variables on USPS→MNIST



VAE CVAE Proposed

Density estimation performance

	VAE	CVAE	Proposed
USPS→MNIST	-163.25 ± 2.15	-152.32 ± 1.64	$\mathbf{-149.08 \pm 0.86}$
MNIST→USPS	-235.23 ± 1.54	$\mathbf{-211.18 \pm 0.55}$	$\mathbf{-212.11 \pm 1.48}$
Synth→SVHN	1146.04 ± 35.65	1397.36 ± 10.89	$\mathbf{1430.27 \pm 11.44}$
SVHN→Synth	760.66 ± 8.85	814.63 ± 10.09	$\mathbf{855.51 \pm 11.41}$

Downstream classification accuracy

	VAE	CVAE	Proposed
USPS→MNIST	0.52 ± 2.15	0.53 ± 0.02	$\mathbf{0.68 \pm 0.01}$
MNIST→USPS	0.64 ± 0.01	0.67 ± 0.01	$\mathbf{0.74 \pm 0.02}$
Synth→SVHN	0.20 ± 0.00	$\mathbf{0.21 \pm 0.00}$	0.19 ± 0.00
SVHN→Synth	0.25 ± 0.01	0.25 ± 0.00	$\mathbf{0.26 \pm 0.00}$