Adversarial Learning of a Variational Generative Model with Succinct Bottleneck Representation

Jongha Jon Ryu[1], Yoojin Choi[2], Young-Han Kim[1], Mostafa El-Khamy[2], Jungwon Lee[2]
[1]University of California, San Diego [2]Samsung Semiconductor Inc.

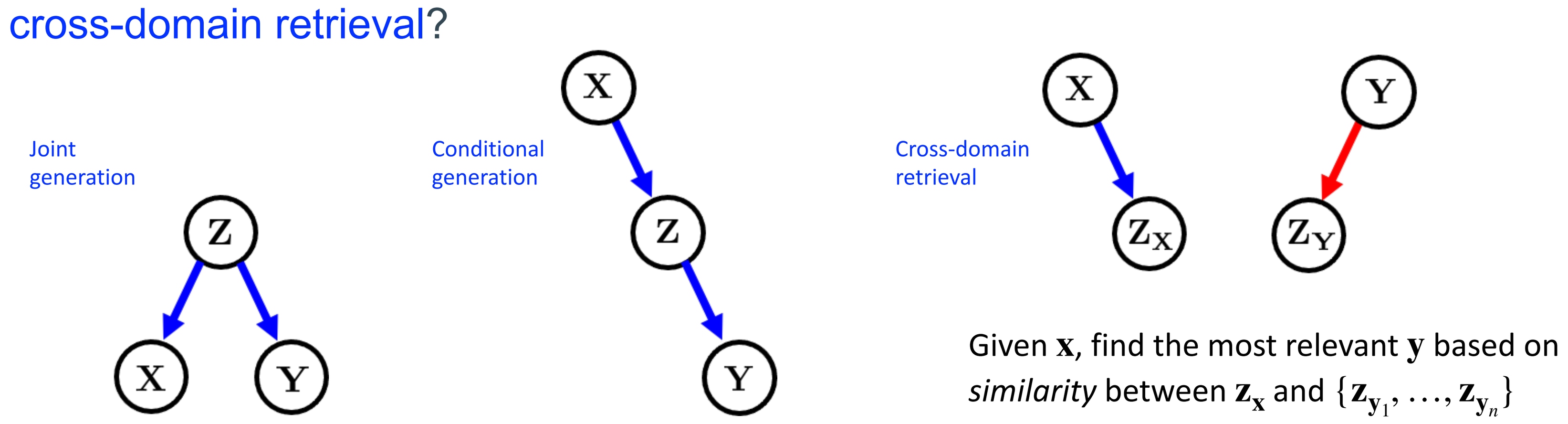
UC San Diego JACOBS SCHOOL OF ENGINEERING Electrical and Computer Engineering

SAMSUNG

NEURAL INFORMATION PROCESSING SYSTEMS

Problem: Cross-Domain Disentanglement

Given two random vectors $\mathbf{X}$ and $\mathbf{Y}$, what is the common representation $\mathbf{Z}$ that is *useful* for downstream tasks such as joint/conditional generation and cross-domain retrieval?



Existing information-theoretic approaches:

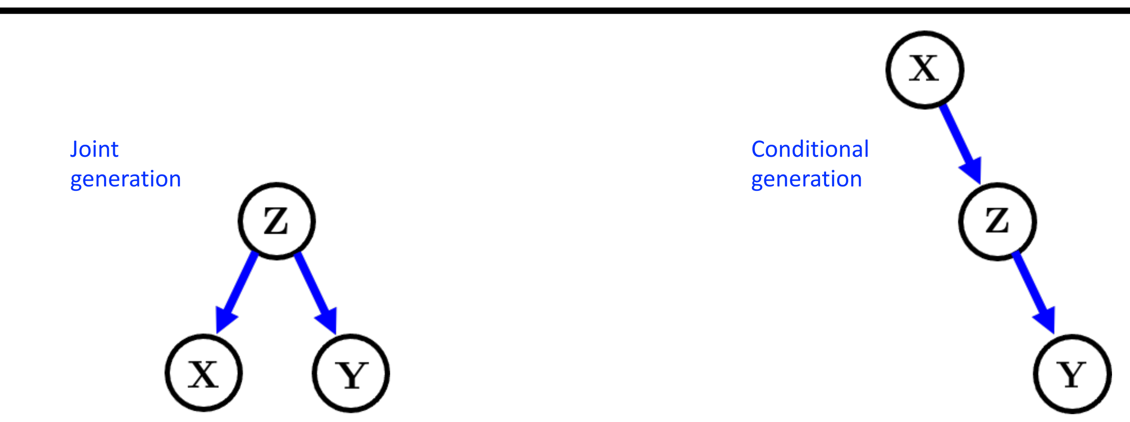
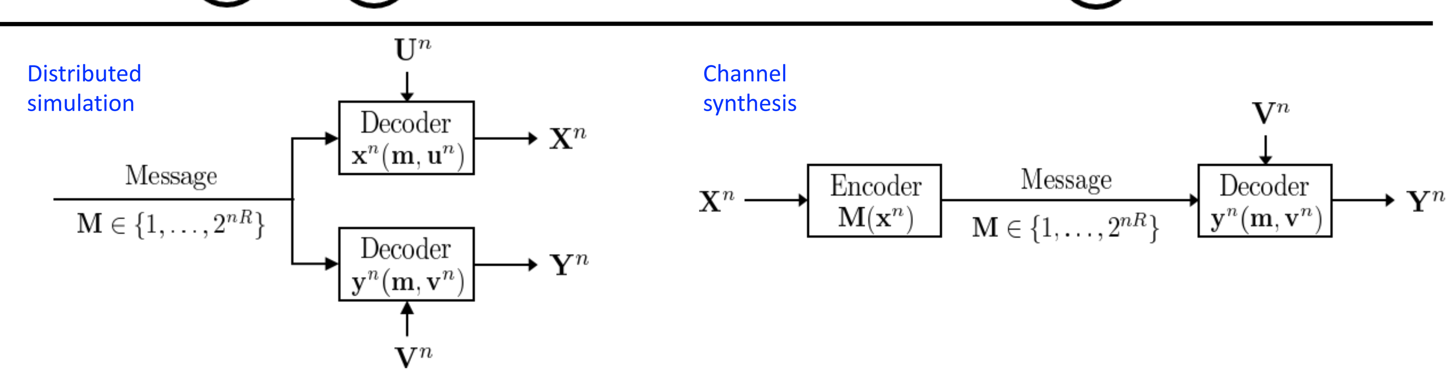
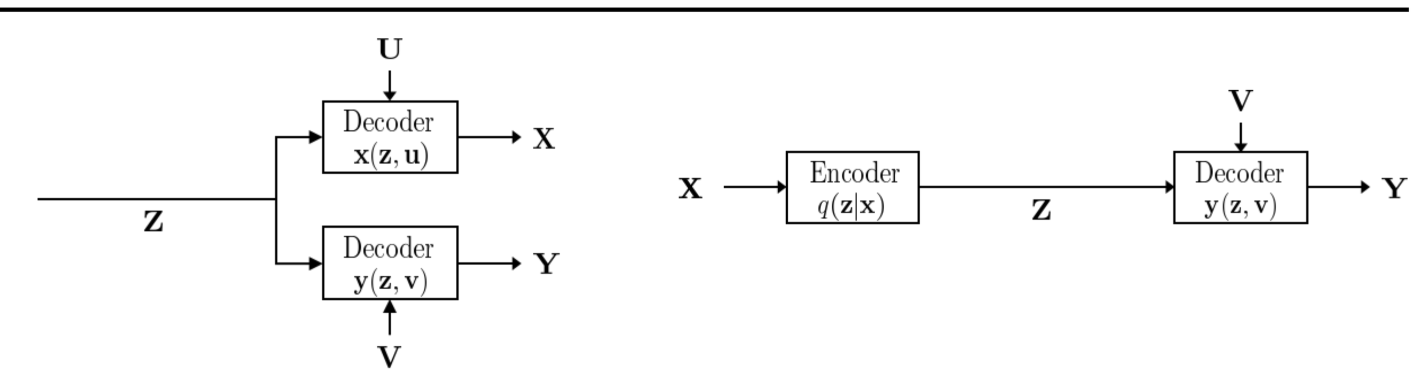
- information bottleneck principle [1],
- interaction information maximization [2], …

They either

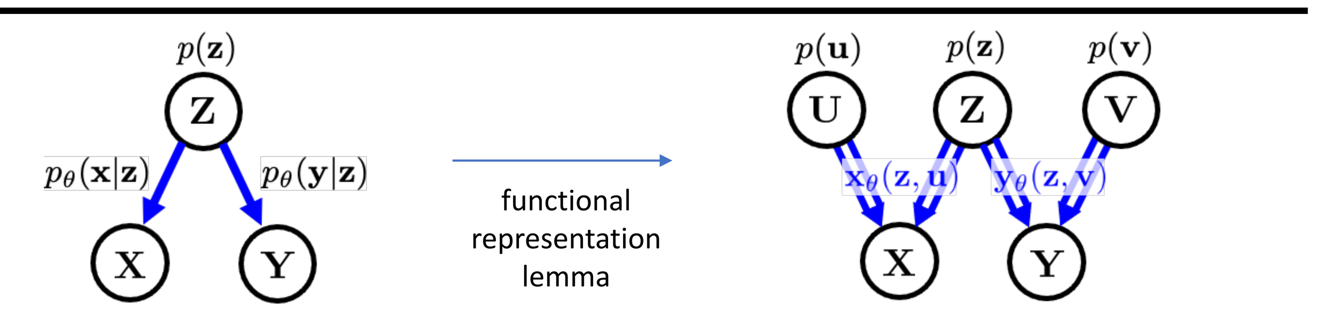
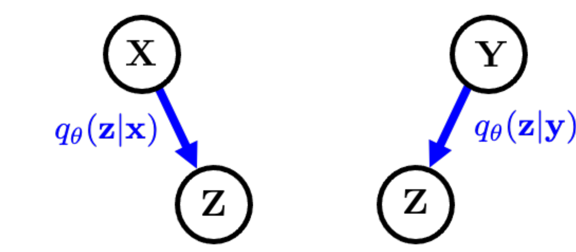
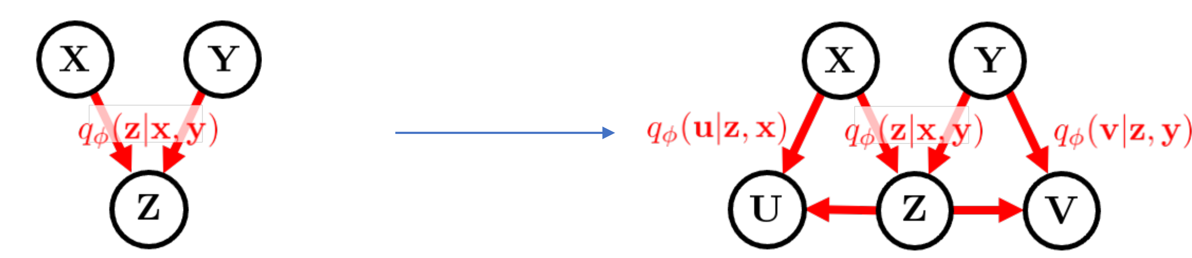
- are not proper to capture many-to-many relations [1] or
- lack a notion of ***optimality*** [1,2]

Wyner's Common Information

as a representation-learning principle

	Joint	Conditional
Generation tasks		
Information-theoretic analogues		
Single-letter characterizations		
Fundamental limit	$J(\mathbf{X};\mathbf{Y}) = \min_{q(\mathbf{z}\mid\mathbf{x},\mathbf{y}):\ \mathbf{X}-\mathbf{Z}-\mathbf{Y}} I(\mathbf{X},\mathbf{Y};\mathbf{Z})$	

The Variational Wyner Model

	Implicit models	Explicit models with local randomness
Decoders		
Encoders		
Variational[1] encoders		

[1]"variational" under matching with f-divergence

Approach: Train them jointly by distribution matching with minimizing $I(\mathbf{X},\mathbf{Y};\mathbf{Z})$!

Training Method

Table 1: Induced distributions and their shorthand notation.

Type	Distribution over $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{v})$	Notation
joint ($\to$ xy)	$p_\theta(\mathbf{z})p_\theta(\mathbf{u})p_\theta(\mathbf{v})\mathbf{x}_\theta(\mathbf{z},\mathbf{u})\mathbf{y}_\theta(\mathbf{z},\mathbf{v})$	$p_{\to\mathsf{xy}}$
cond. (x $\to$ y)	$q(\mathbf{x})q_\theta(\mathbf{z}\mid\mathbf{x})p_\theta(\mathbf{v})\mathbf{y}_\theta(\mathbf{z},\mathbf{v})q_\phi(\mathbf{u}\mid\mathbf{z},\mathbf{x})$	$p_{\mathsf{x}\to\mathsf{y}}$
cond. (y $\to$ x)	$q(\mathbf{y})q_\theta(\mathbf{z}\mid\mathbf{y})p_\theta(\mathbf{u})\mathbf{x}_\theta(\mathbf{z},\mathbf{u})q_\phi(\mathbf{v}\mid\mathbf{z},\mathbf{y})$	$p_{\mathsf{y}\to\mathsf{x}}$
variational (xy $\to$)	$q(\mathbf{x},\mathbf{y})q_\phi(\mathbf{z}\mid\mathbf{x},\mathbf{y})q_\phi(\mathbf{u}\mid\mathbf{z},\mathbf{x})q_\phi(\mathbf{v}\mid\mathbf{z},\mathbf{y})$	$q_{\mathsf{xy}\to}$

Training objective

- matching distributions with symmetric KL divergence
- common information regularization

$$\text{minimize} \quad \mathcal{D}^{\mathsf{xyzuv}}_{\mathsf{model}} + \lambda^{\mathsf{CI}}_{\mathsf{model}} I_{\mathsf{model}}(\mathbf{X},\mathbf{Y};\mathbf{Z})$$

- + reconstruction consistency; latent matching; cross matching

Approximate training with adversarial density ratio estimation

Variation of "GAN training"

Plug in and maximize

$$D_{\mathsf{JS}}(p(\mathbf{s}), q(\mathbf{s})) = \max_{r(\mathbf{s})} \{\mathbb{E}_{p(\mathbf{s})}[\log\sigma(\log r(\mathbf{s}))] + \mathbb{E}_{q(\mathbf{s})}[\log\sigma(-\log r(\mathbf{s}))]\}$$

$$D_{\mathsf{sym}}(p(\mathbf{s}), q(\mathbf{s})) \approx \mathbb{E}_{p(\mathbf{s})}[\log r(\mathbf{s})] - \mathbb{E}_{q(\mathbf{s})}[\log r(\mathbf{s})]$$

Plug in $r(\mathbf{s})$ and minimize

Training details

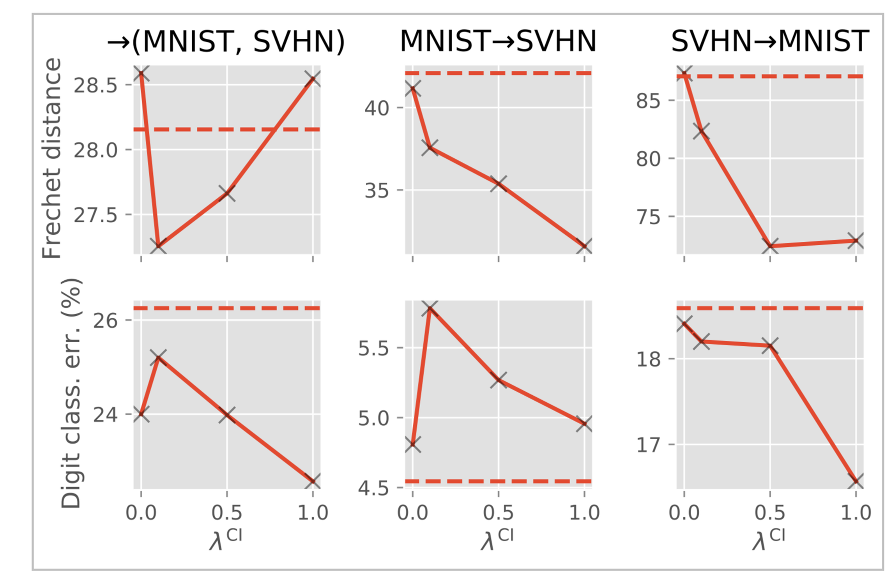
- Shared discriminator feature maps (for efficiency)
- Deterministic encoders (for unbiased gradient estimation)

Experiments

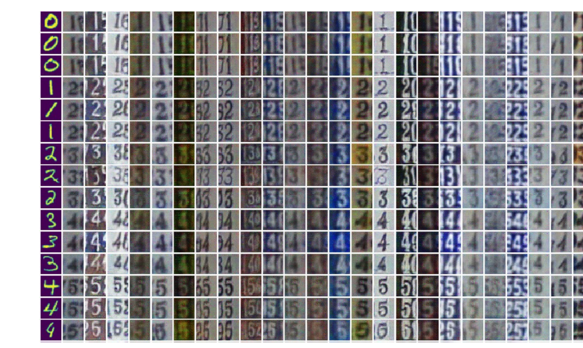
(1) MNIST-SVHN add-1 dataset

Joint/conditional generation, latent classification

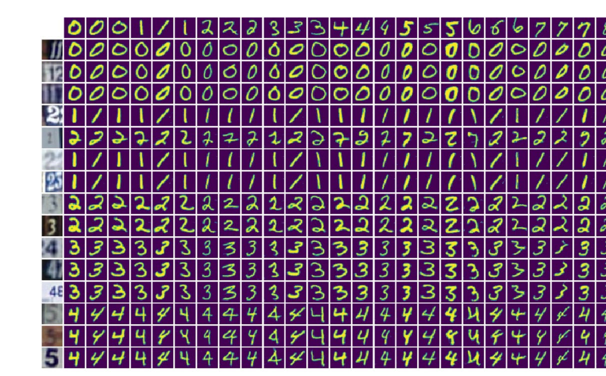
Increasing CI regularization (λ^{CI}) improves sample quality and representation quality (with tradeoff)!



(a) Joint generation



(b) Conditional generation

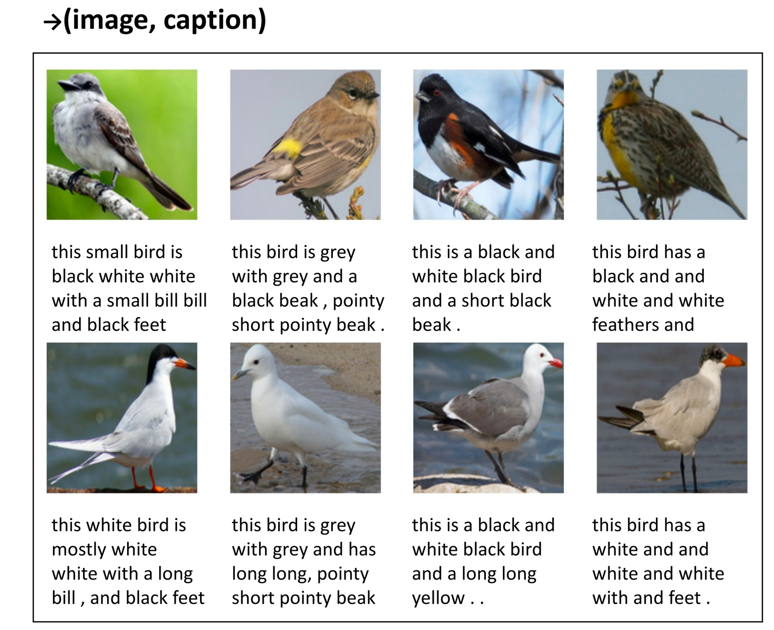


(c) Conditional generation with style transfer

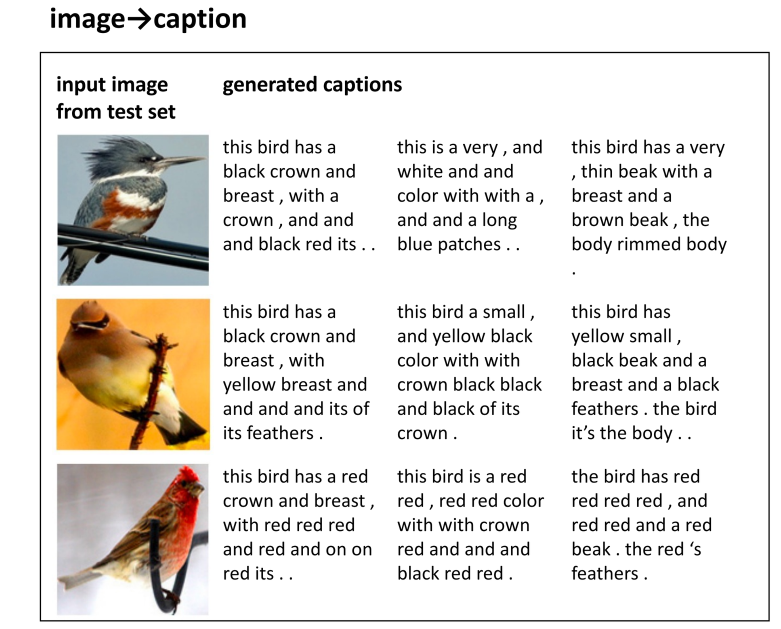
(2) CUB image-caption dataset

Caption ↔ Image feature

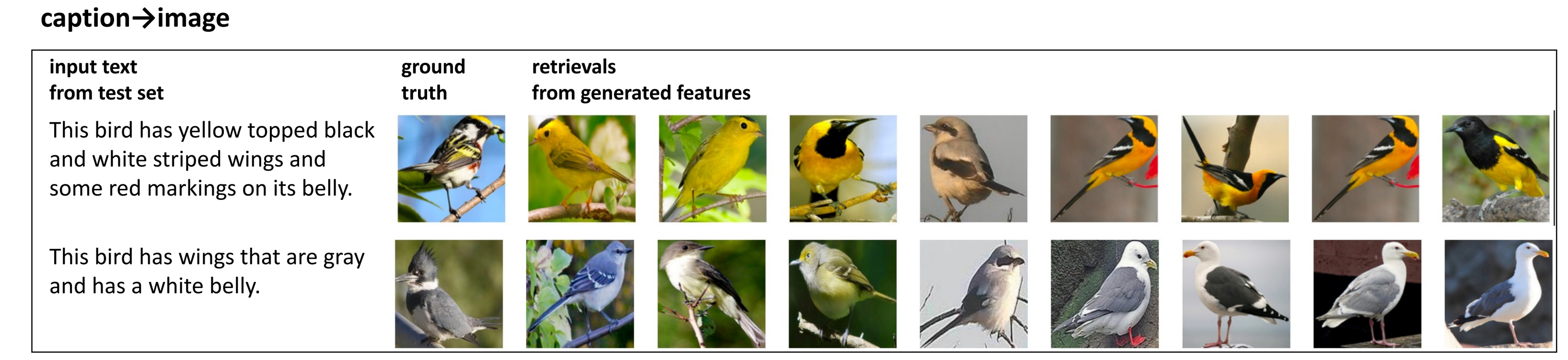
→(image, caption)



image→caption



caption→image

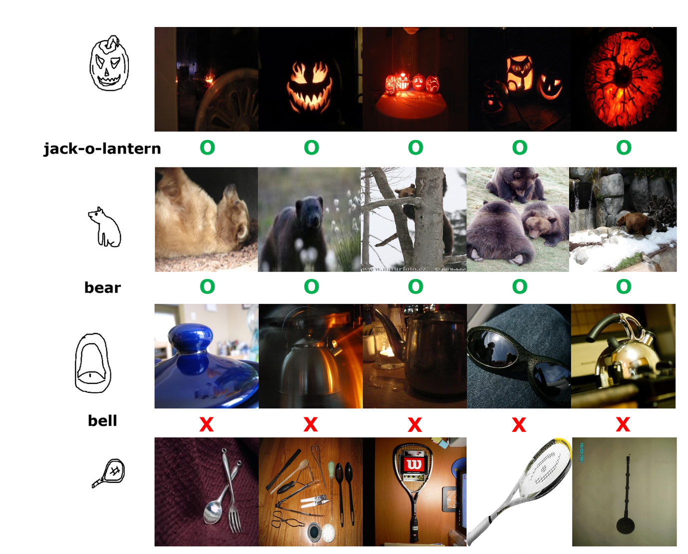


(3) Sketchy dataset

Zero-shot sketch-based image retrieval

Table 2: Evaluation of the ZS-SBIR task with the Sketchy Extended dataset.

Models	P@100	mAP
LCALE [11]	0.583	0.476
IIAE [6]	0.659	0.573
Variational Wyner	**0.703**	**0.629**



References

[1] N. Tishby et al. (1999) "The information bottleneck method". In: *Proc. 37th Annual Allerton Conference on Communications, Control and Computing*, pp. 368–377.
[2] H. Hwang et al. (2020). "Variational Interaction Information Maximization for Cross-domain Disentanglement". In: *Proc. Advances in Neural Information Processing Systems, 33.*