Variational Interaction Information Maximization for Cross-domain Disentanglement

Reporter：陈思玉

2024.3.23

Hwang H J, Kim G H, Hong S, et al. Variational interaction information maximization for cross-domain disentanglement[J]. Advances in Neural Information Processing Systems, 2020, 33: 22479-22491.

Author

HyeongJoo Hwang

KAIST

在 ai.kaist.ac.kr 的电子邮件经过验证

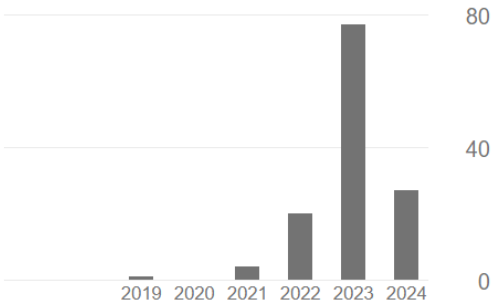
关注

创建我的个人资料

标题	引用次数	年份
Demodice: Offline imitation learning with supplementary imperfect demonstrations CCF none GH Kim, S Seo, J Lee, W Jeon, HJ Hwang, H Yang, KE Kim International Conference on Learning Representations	60	2021
Variational interaction information maximization for cross-domain disentanglement CCF A HJ Hwang, GH Kim, S Hong, KE Kim Advances in Neural Information Processing Systems 33, 22479-22491	40	2020
Multi-view representation learning via total correlation objective CCF A HJ Hwang, GH Kim, S Hong, KE Kim Advances in Neural Information Processing Systems 34, 12194-12207	29	2021
Regularized Behavior Cloning for Blocking the Leakage of Past Action Information CCF A S Seo, HJ Hwang, H Yang, KE Kim Advances in Neural Information Processing Systems 36		2024
Information-theoretic state space model for multi-view reinforcement learning CCF A HJ Hwang, S Seo, Y Jang, S Kim, GH Kim, S Hong, KE Kim		2023
사전확률 갱신을 수행하는 교차 엔트로피 계획법 CCF none 황형주, 장영수, 박재영, 김기응 정보과학회논문지 47 (1), 88-94		2020

引用次数

	总计	2019 年至今
引用	129	129
h 指数	3	3
i10 指数	3	3



Cross-Domain Disentanglement

定义

跨域解耦（Cross-Domain Disentanglement）从两个域中将表征划分为：

- 域不变（Domain-Invariant）表征
- 域特定（Domain-Specific）表征

举例

MNIST-CDCB 数据集

- 域不变表征：数字
- 域特定表征：前景、背景

Interaction Information Auto-Encoder

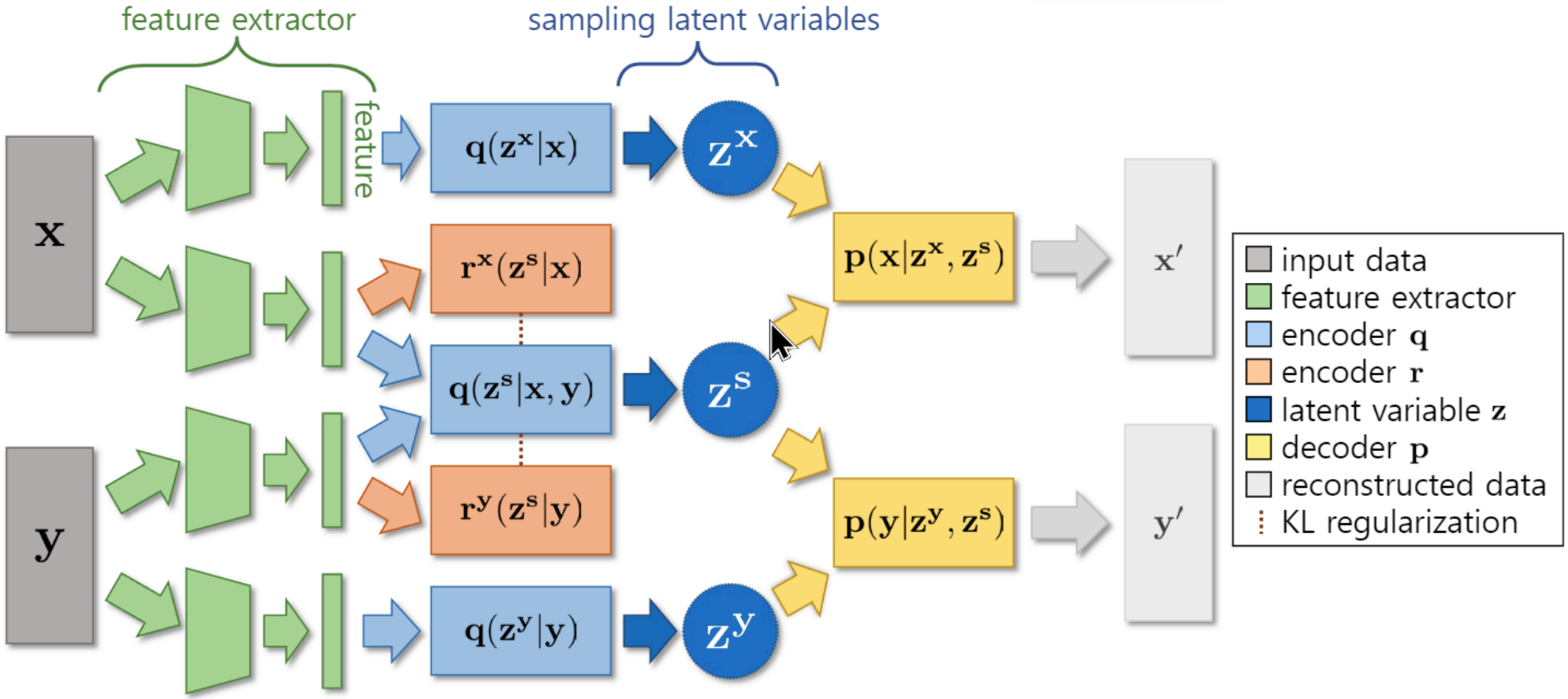


Figure 2: The architecture of Interaction Information Auto-Encoder.

Formulation

未知联合分布 $(x, y) \sim p_D(x, y)$，其中 x 和 y 来自不同的域 X 和 Y。

两个域存在特定于域的因素，同时存在一些共同的因素。
例如，x 和 y 可以是相同内容但风格不同的（例如草图和照片）的图像

问题：从给定的数据中解耦出三个部分：域特定特征 Z^X, Z^Y 和域共享特征 Z^S

(a) Generative model p_θ (b) Approximate inference model q_ϕ

Figure 1: Graphical models for cross-domain disentanglement.

Formulation

Goal

$$p_\theta(x,y) = \int dz^x dz^s dz^y p_{\theta_X}(x|z^x,z^s) p_{\theta_Y}(y|z^y,z^s) p(z^x)p(z^s)p(z^y)$$

$$\Rightarrow \quad q_\phi(z^x,z^s,z^y|x,y) \quad \underrightarrow{\text{近似}} \quad p_\theta(z^x,z^s,z^y|x,y)$$

$$\begin{aligned}
\Rightarrow \log p(x,y) \quad &\geq \mathbb{E}_{q(z^x,z^s,z^y|x,y)}\left[\log \frac{p(x,y,z^x,z^s,z^y)}{q(z^x,z^s,z^y|x,y)}\right] \\
&= \mathbb{E}_{q(z^x|x)q(z^s|x,y)}[\log p(x|z^x,z^s)] + \mathbb{E}_{q(z^y|y)q(z^s|x,y)}[\log p(y|z^y,z^s)] \\
&\quad - D_{KL}[q(z^x|x)\|p(z^x)] - D_{KL}[q(z^y|y)\|p(z^y)] \\
&\quad - D_{KL}[q(z^s|x,y)\|p(z^s)]
\end{aligned}$$

不能满足：

- Z^S 与 Z^X、Z^Y 互斥
- Z^X, Z^Y 只包含域特定特征，Z^S 只包含域共享特征

Formulation

保证 Z^S 与 Z^X、Z^Y 互斥

$$\min I(Z^X; Z^S) = -I(X; Z^X, Z^S) + I(X; Z^X) + I(X; Z^S)$$

保证 Z^X, Z^Y 只包含域特定特征，Z^S 只包含域共享特征

$$\begin{aligned}\max I(X; Y; Z^S) &= I(X; Z^S) - I(X; Z^S|Y) \\ &= I(Y; Z^S) - I(Y; Z^S|X)\end{aligned}$$

联合优化

$$\begin{aligned}&\max_q I(X; Y; Z^S) - I(Z^X; Z^S) \\ &= \underbrace{\cancel{I(X; Z^S)} - I(X; Z^S|Y)}_{I(X;Y;Z^S)} + \underbrace{I(X; Z^X, Z^S) - I(X; Z^X) - \cancel{I(X; Z^S)}}_{-I(Z^X;Z^S)} \\ &= I(X; Z^X, Z^S) - I(X; Z^X) - I(X; Z^S|Y) \qquad (8)\end{aligned}$$

Formulation

$$I(X; Z^X, Z^S)$$

$$q(x|z^x, z^s) = \frac{q(z^x, z^s|x)\underbrace{p_D(x)}}{\int \underbrace{p_D(x, y)}q(z^x, z^s|x, y)dxdy} \quad \Rightarrow \quad \text{难以计算}$$

$$\Rightarrow p(x|z^x, z^s) \quad \underrightarrow{\text{计算}} \quad q(x|z^x, z^s)$$

$$\begin{aligned}
& I(X; Z^X, Z^S) \\
&= \mathbb{E}_{q(z^x, z^s|x)p_D(x)}\left[\log \frac{q(x|z^x, z^s)}{p_D(x)}\right] \\
&= H(X) + \mathbb{E}_{q(z^x, z^s|x)p_D(x)}[\log p(x|z^x, z^s)] + \mathbb{E}_{q(z^x, z^s)}[D_{KL}[q(x|z^x, z^s)\|p(x|z^x, z^s)]] \\
&\geq H(X) + \mathbb{E}_{q(z^x, z^s|x)p_D(x)}[\log p(x|z^x, z^s)] \\
&= H(X) + \mathbb{E}_{p_D(x,y)q(z^x|x)q(z^s|x,y)}[\log p(x|z^x, z^s)]
\end{aligned}$$

Formulation

$-I(X;Z^X)$

$$q(z^x) = \int \underbrace{p_D(x)} q(z^s|x)dx \quad \Rightarrow \quad \text{难以计算}$$

$$\Rightarrow \ -\mathbb{E}_{p_D(x)}[D_{KL}[q(z^x|x)\|p(z^x)]]$$

Formulation

$$-I(X;Z^S|Y)$$

$$q(z^s|y)=\int \underbrace{p_D(x|y)}q(z^s|x,y)dx \quad \Rightarrow \quad \text{难以计算}$$

$$\Rightarrow r^y(z^s|y) \quad \underrightarrow{\text{计算}} \quad q(z^s|y)$$

$$\begin{aligned}
& -I(X;Z^S|Y) \\
= & -\mathbb{E}_{p_D(x,y)q(z^s|x,y)}\left[\log\frac{q(z^s|x,y)}{q(z^s|y)}\right] \\
= & -\mathbb{E}_{p_D(x,y)q(z^s|x,y)}\left[\log\frac{q(z^s|x,y)r^y(z^s|y)}{r^y(z^s|y)q(z^s|y)}\right] \\
= & -\mathbb{E}_{p_D(x,y)}[D_{KL}[q(z^s|x,y)\|r^y(z^s|y)]]+\mathbb{E}_{p_D(y)}[D_{KL}[q(z^s|y)\|r^y(z^s|y)]] \\
\geq & -\mathbb{E}_{p_D(x,y)}[D_{KL}[q(z^s|x,y)\|r^y(z^s|y)]]
\end{aligned}$$

Formulation

最终目标

$$
\begin{aligned}
&(I(X;Y;Z^S) - I(Z^X;Z^S)) + (I(X;Y;Z^S) - I(Z^Y;Z^S)) \\
&= 2 \cdot I(X;Y;Z^S) - I(Z^X;Z^S) - I(Z^Y;Z^S) \\
&= I(X;Z^X,Z^S) + I(Y;Z^Y,Z^S) - I(X;Z^X) - I(Y;Z^Y) - I(X;Z^S|Y) - I(Y;Z^S|X) \\
&\geq \mathbb{E}_{p_D(x,y)}\left[\mathbb{E}_{q(z^s|x,y)q(x^x|x)}[\log p(x|z^x,z^s)] + \mathbb{E}_{q(z^s|x,y)q(z^y|y)}[\log p(y|z^y,z^s)]\right] \\
&\quad - \mathbb{E}_{p_D(x,y)}[D_{KL}[q(z^x|x)\|p(z^x)] + D_{KL}[q(z^y|y)\|p(z^y)]] \\
&\quad - \mathbb{E}_{p_D(x,y)}[D_{KL}[q(z^s|x,y)\|r^y(z^s|y)] + D_{KL}[q(z^s|x,y)\|r^x(z^s|x)]] \\
&\quad + H(X) + H(Y)
\end{aligned}
$$

Interaction Information Auto-Encoder

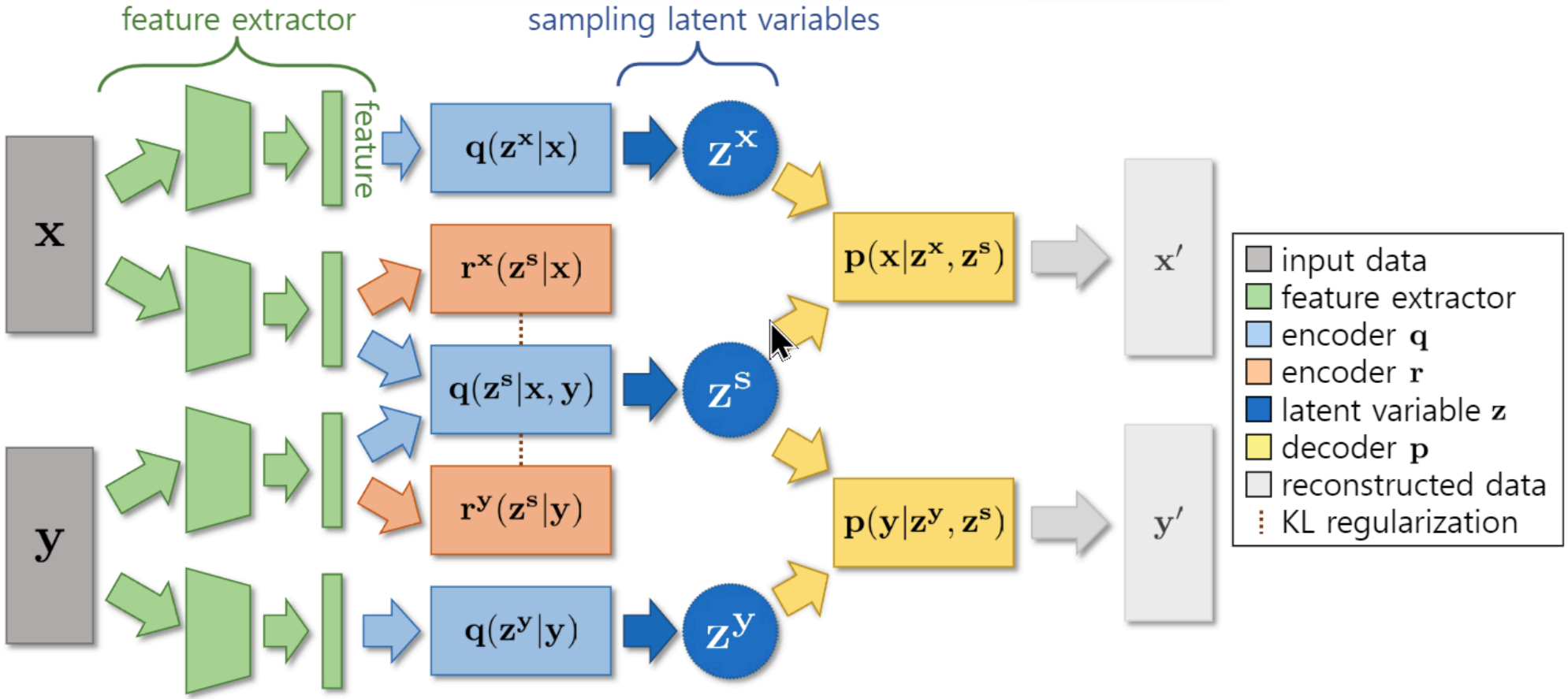


Figure 2: The architecture of Interaction Information Auto-Encoder.

Loss

$$\max_{p,q} \mathbb{E}_{q(z^x,z^s,z^y,x,y)} \left[\log \frac{p(x,y,z^x,z^s,z^y)}{q(z^x,z^s,z^y|x,y)} \right] + \lambda(2 \cdot I(X;Y;Z^S) - I(Z^X;Z^S) - I(Z^Y;Z^S))$$

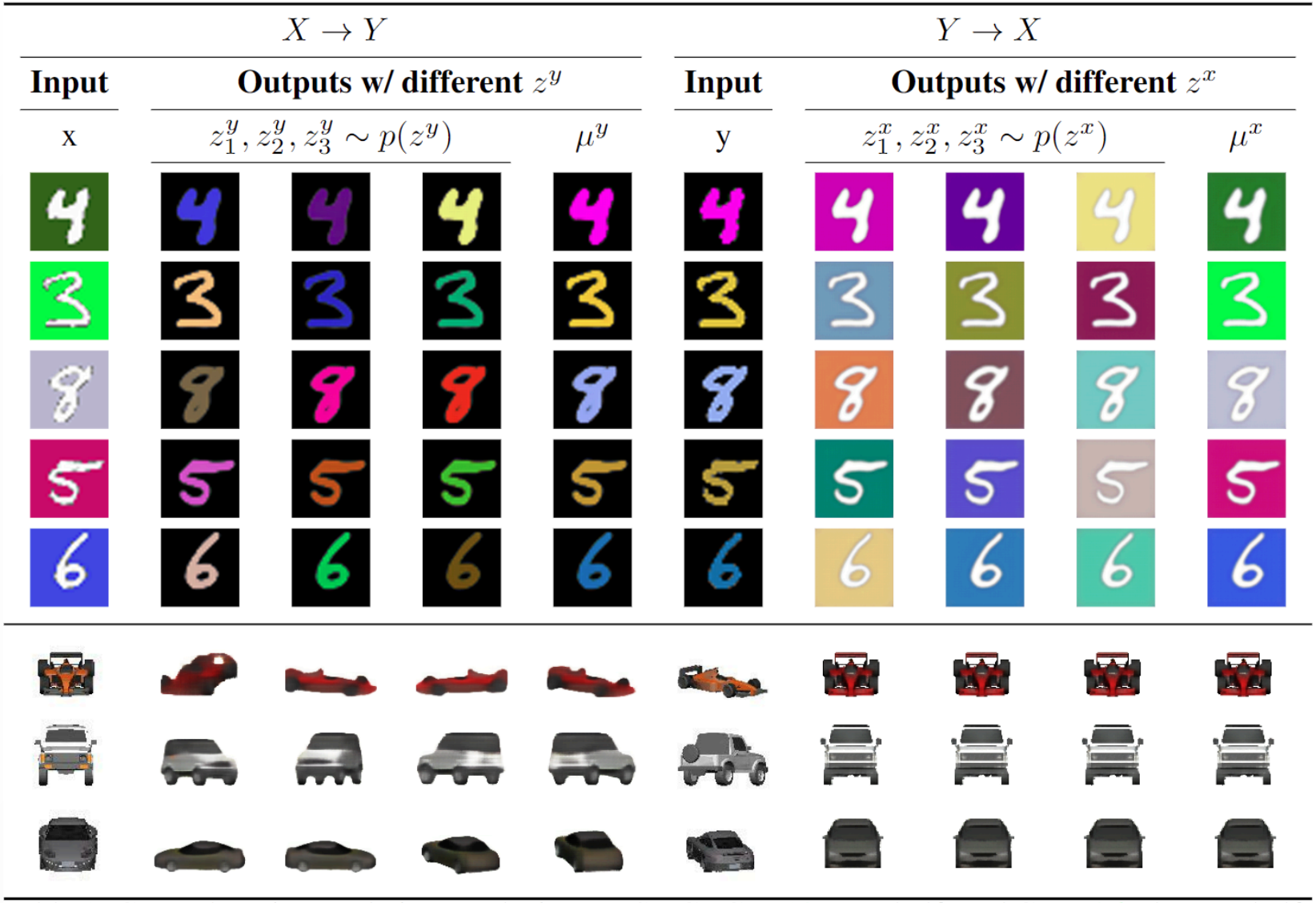
$$\geq \max_{p,q,r} (1+\lambda) \cdot \mathbb{E}_{p_D(x,y)}[ELBO(p,q)]$$

$$+ \lambda \cdot \mathbb{E}_{p_D(x,y)}[D_{KL}[q(z^s|x,y) \| p(z^s)]]$$

$$- \lambda \cdot \mathbb{E}_{p_D(x,y)}[D_{KL}[q(z^s|x,y) \| r^y(z^s|y)] + D_{KL}[q(z^s|x,y) \| r^x(z^s|x)]]$$

Experiments

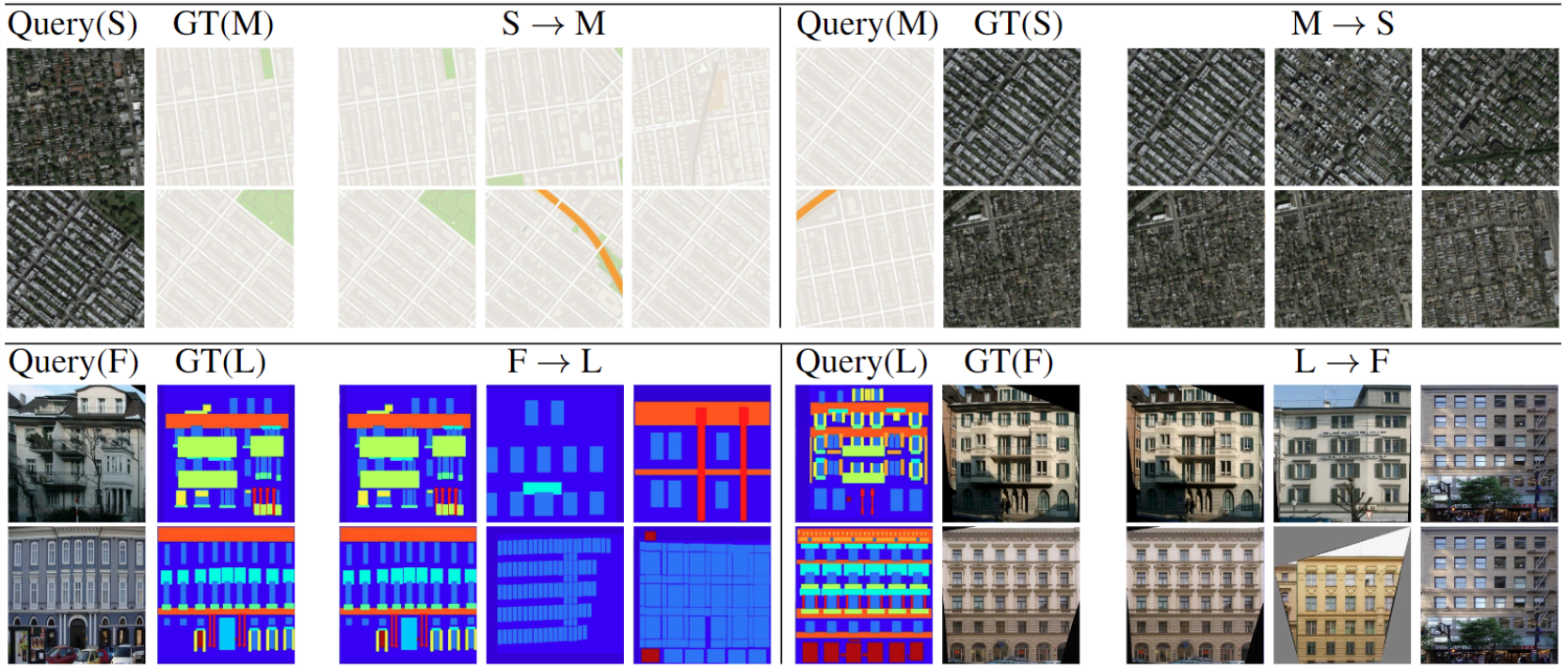
Cross-domain Image Translation



Cross-Domain retrieval

Table 2: Shared (exclusive) representation based retrieval on MNIST-CDCB [12], Maps [17], and Facades [41] dataset. CD/CB stand for colored digit/background, S/M stand for satellite/map, and F/L stand for facade/label respectively.

Dataset	**MNIST-CDCB**		**Maps**		**Facades**	
Models	CD → CB	CB → CD	S → M	M → S	F → L	L → F
DRIT [26]	-	-	33.8 (0.09)	37.3 (0.09)	31.1 (0.94)	44.3 (0.94)
CdDN [12]	99.6 (0.0)	99.6 (0.0)	91.4 (0.18)	96.9 (0.09)	84.9 (0.94)	89.6 (0.0)
IIAE	**99.7** (0.01)	**99.7** (0.01)	**96.6** (0.09)	**97.3** (0.0)	**96.2** (0.94)	**99.1** (0.94)



Experiments

Zero-shot Sketch based Image Retrieval

Table 3: Evaluation on the Sketchy Extended dataset [29, 37]. WordEmb stands for word embedding.

Models	Feature Dimension	Evaluation metric mAP	Evaluation metric P@100	External knowledge Attribute	External knowledge WordEmb.	External knowledge WordNet [33]
SAE [23]	300	0.216	0.293	✓	✓	-
FRWGAN [9]	512	0.127	0.169	✓	-	-
ZSIH [38]	64	0.258	0.342	-	✓	-
CAAE [22]	4096	0.196	0.284	-	-	-
SEM-PCYC [6]	64	0.349	0.463	-	✓	✓
LCALE [27]	64	0.476	0.583	-	✓	-
IIAE	64	**0.573**	**0.659**	-	-	-

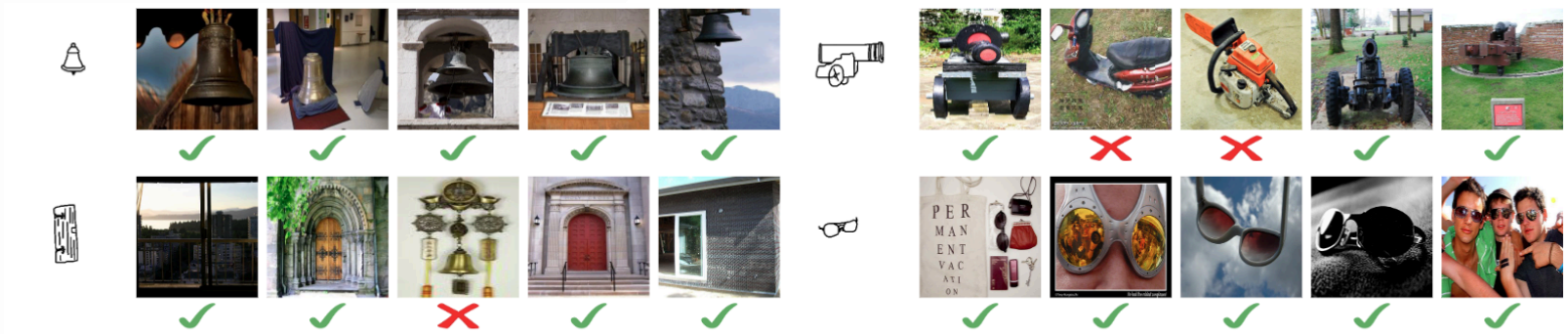


Figure 4: Top-5 ZS-SBIR samples from IIAE on the Sketchy Extended dataset. Sketches in the first and seventh columns are queries and rest are retrieved candidates (Top-1 to 5 from the left to the right). Green checkmark indicates correct retrieval, whereas red crossmark indicates wrong retrieval.