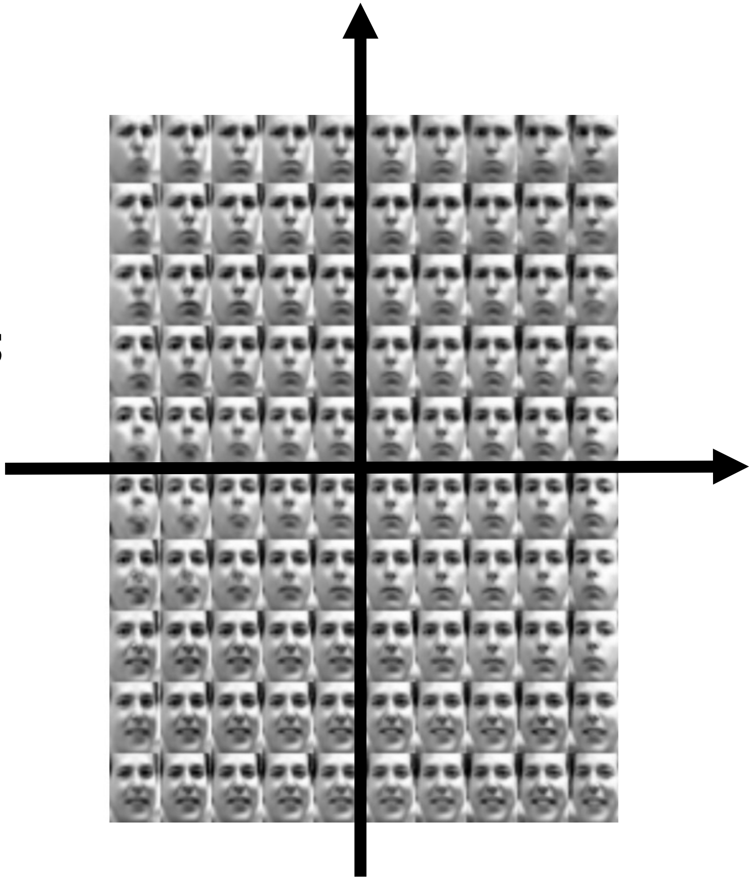
# Variational Interaction Information Maximization for Cross-domain Disentanglement

Accepted in NeurIPs 2020

Kyeong Ryeol, Go
M.S. Candidate of OSI Lab

# Disentanglement

- What for?
  - Identifying sources of variation for interpretability
  - Obtaining representation invariant to nuisance factors
  - Domain transfer

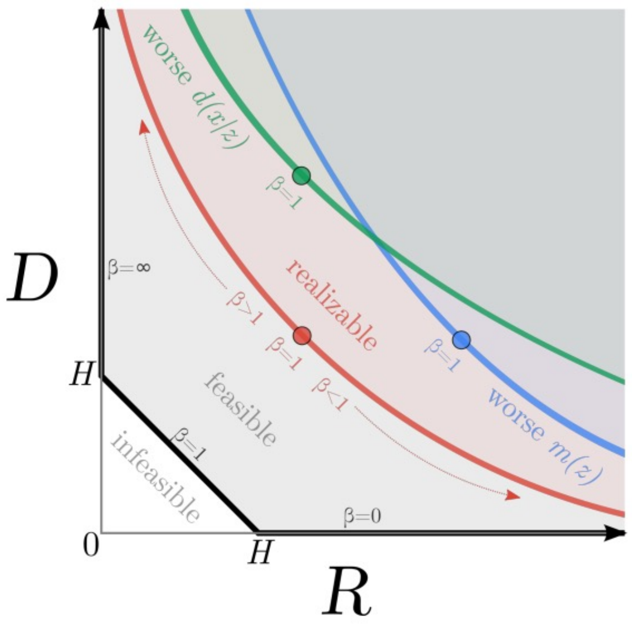


- Variational Autoencoder (Arxiv 2014)
  - $\log p(x) \geq \underbrace{E_{q_\phi(z|x)}[\log p_\theta(x|z)]}_{\text{Reconstruction term}} - \underbrace{KL(q_\phi(z|x)\|p(z))}_{\text{Compression term}}$

- VAE failures
  - Amortized Inference failures
    - ELBO can be maximized even with inaccurate variational posterior
    - Error in $x$ is more critical than in $z$ due to high dimensionality → overfitting
  - Information preference property
    - Complex decoder improves sample quality while neglecting the latent variable

# Disentanglement

- ELBO variants
  - $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot KL(q_\phi(z|x)||p(z))$
  - $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot KL\left(q_\phi(z) \middle\| \prod_d q_\phi(z_d)\right) - KL(q_\phi(z|x)||p(z))$
  - $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot HSIC\left(q_\phi(z)\right) - KL\left(q_\phi(z|x)\|p(z)\right)$
  - …
- Information theory
  - $H - D \le I_q(x,z) = KL\left(q_\phi(z|x)p(x)\middle\|q_\phi(z)p(x)\right) \le R$
  - Data entropy : $H = -\int p(x)\log p(x)\,dx$
  - Distortion : $D = -E_{p(x)q_\phi(z|x)}[\log p_\theta(x|z)]$
  - Rate : $R = E_{p(x)}\left[KL\left(q_\phi(z|x)\middle\|\mathrm{r}_\psi(z)\right)\right]$
  - $\max\limits_{\theta,\phi,\psi} -D - |R - \sigma|$



# Cross domain disentanglement

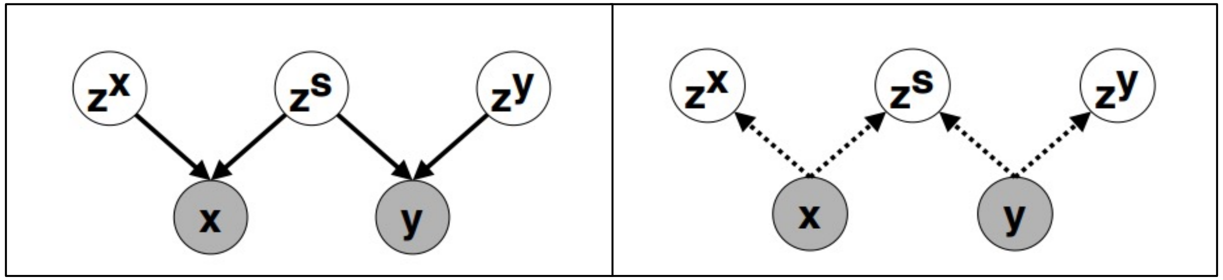
- What for?
  - Successful domain transfer
  - Measuring semantic distance between domains

- Goal
  - Partitioning into <u>domain-invariant($z^s$)</u> and <u>domain specific($z^x, z^y$)</u>

    Ex) Images in different styles with similar semantic content

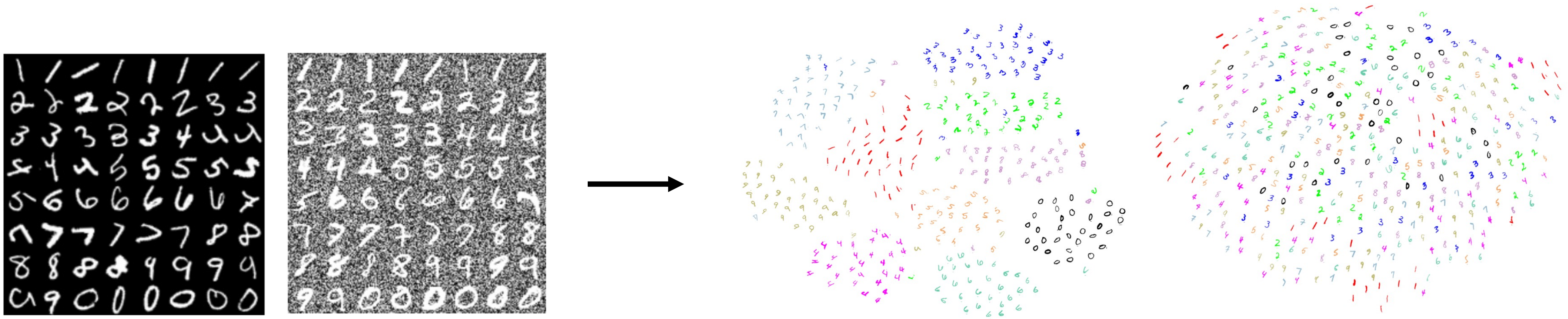
    1) Maximizes the joint distribution $p_D(x, y)$ by optimizing $\theta$

    2) Disentangle $z^x, z^y$ from $z^s$



# Baseline

- Deep Variational Canonical Correlation Analysis (Arxiv 2016)
  - $p_\theta(x, y, z^x, z^s, z^y) = p_{\theta_X}(x|z^x, z^s)p_{\theta_Y}(y|z^y, z^s)p(z^x)p(z^s)p(z^y)$
  - $q_\phi(z^x, z^s, z^y|x, y) = q_{\phi_X}(z^x|x)q_{\phi_S}(z^s|x, y)q_{\phi_Y}(z^y|y)$
  - $L_0 \coloneqq E_{q_\phi(z^x, z^s, z^y|x, y)}\left[\log \frac{p_\theta(x,y,z^x,z^s,z^y)}{q_\phi(z^x, z^s, z^y|x, y)}\right]$
    $= E_{q_{\phi_X}(z^x|x)q_{\phi_S}(z^s|x,y)}[\log p_\theta(x|z^x, z^s)] + E_{q_{\phi_Y}(z^y|y)q_{\phi_S}(z^s|x, y)}[\log p_\theta(y|z^y, z^s)]$
    $- KL\big(q_{\phi_X}(z^x|x)\|p(z^x)\big) - KL\big(q_{\phi_S}(z^s|x, y)\|p(z^s)\big) - KL\big(q_{\phi_Y}(z^y|y)\|p(z^y)\big)$



$x$ $y$ $z^s$ $z^x$

# Main idea

- No control over the assignment of generative factors!
- Devise additional regularization term
  - **maximize** $I(x,y,z^s) = I(x,z^s) - I(x,z^s|y) = I(y,z^s) - I(y,z^s|x)$
  - **minimize** $I(z^x,z^s) = I(x,z^x) + I(x,z^s) - I(x,\{z^x,z^s\})$
  - **minimize** $I(z^y,z^s) = I(y,z^y) + I(y,z^s) - I(y,\{z^y,z^s\})$

* $L_0 + \beta \cdot [\boxed{I(x,y,z^s) - I(z^x,z^s)} + \boxed{I(x,y,z^s) - I(z^y,z^s)}]$

$\geq L_0 + \beta \cdot \Big(L_0 + KL\big(q_{\phi_S}(z^s|x,y)\|p(z^s)\big) - KL\big(q_{\phi_S}(z^s|x,y)\|r_\psi(z^s|x)\big) - KL\big(q_{\phi_S}(z^s|x,y)\|r_\psi(z^s|y)\big)\Big) + \mathrm{C}$

※ In case of $x$…

- $I(x,y,z^s) - I(z^x,z^s) = \boxed{I(x,\{z^x,z^s\})} - \boxed{I(x,z^x)} - \boxed{I(x,z^s|y)}$

1. $I(x,\{z^x,z^s\}) = E_{q_\phi(z^x,z^s|x,y)p(x)}\left[\log \frac{q_\phi(x|z^x,z^s)}{p(x)}\right] \geq H(x) + E_{q_\phi(z^x,z^s|x,y)p(x)}[\log p_\theta(x|z^x,z^s)]$
2. $I(x,z^x) = E_{q_{\phi_X}(z^x|x)p(x)}\left[\log \frac{q_{\phi_X}(z^x|x)}{q_{\phi_X}(z^x)}\right] \leq E_{p(x)}\big[KL\big(q_{\phi_X}(z^x|x)\|p(z^x)\big)\big]$
3. $I(x,z^s|y) = E_{q_{\phi_S}(z^s|x,y)p(x,y)}\left[\log \frac{q_{\phi_S}(z^s|x,y)}{q_{\phi_S}(z^s|y)}\right] \leq E_{p(x,y)}\big[KL\big(q_{\phi_S}(z^s|x,y)\|r_\psi(z^s|y)\big)\big]$

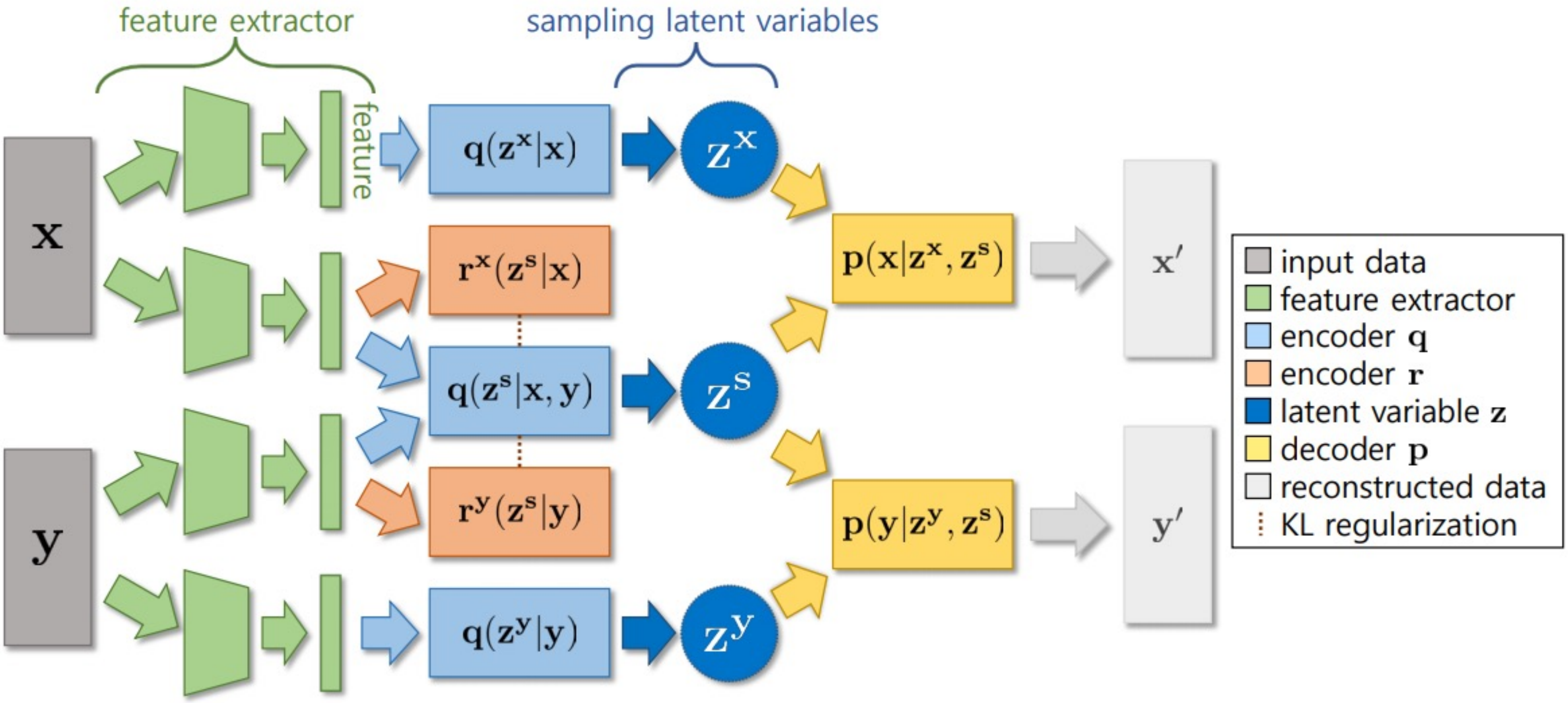



Figure 2: The architecture of Interaction Information Auto-Encoder.

# Experiment

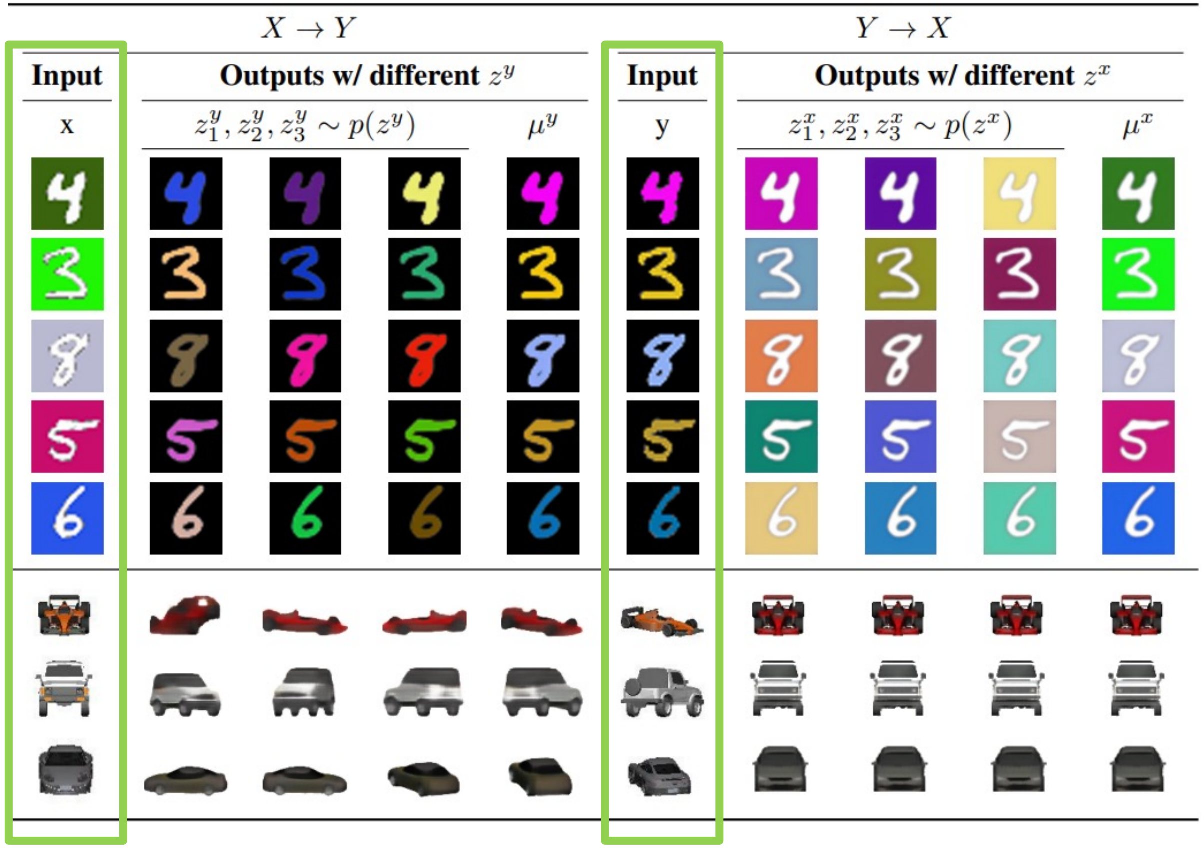
- Cross domain Image translation
  - MNIST-CDCB
    - X : variations in the background
    - Y : variations in the foreground
  - Cars
    - X : frontal view
    - Y : rotated view

※ $X \rightarrow Y$

1. $\mu_x^s = r_\psi(z^s|x).mean$
2. $z^y \sim p(z^y)$ or $z^y = q_{\phi_Y}(z^y|y).mean$
3. $y' = p_\theta(y|\mu_x^s, z^y).mean$

※ $Y \rightarrow X$

1. $\mu_y^s = r_\psi(z^s|y).mean$
2. $z^x \sim p(z^x)$ or $z^x = q_{\phi_X}(z^x|x).mean$
3. $x' = p_\theta(x|\mu_y^s, z^x).mean$



# Experiment

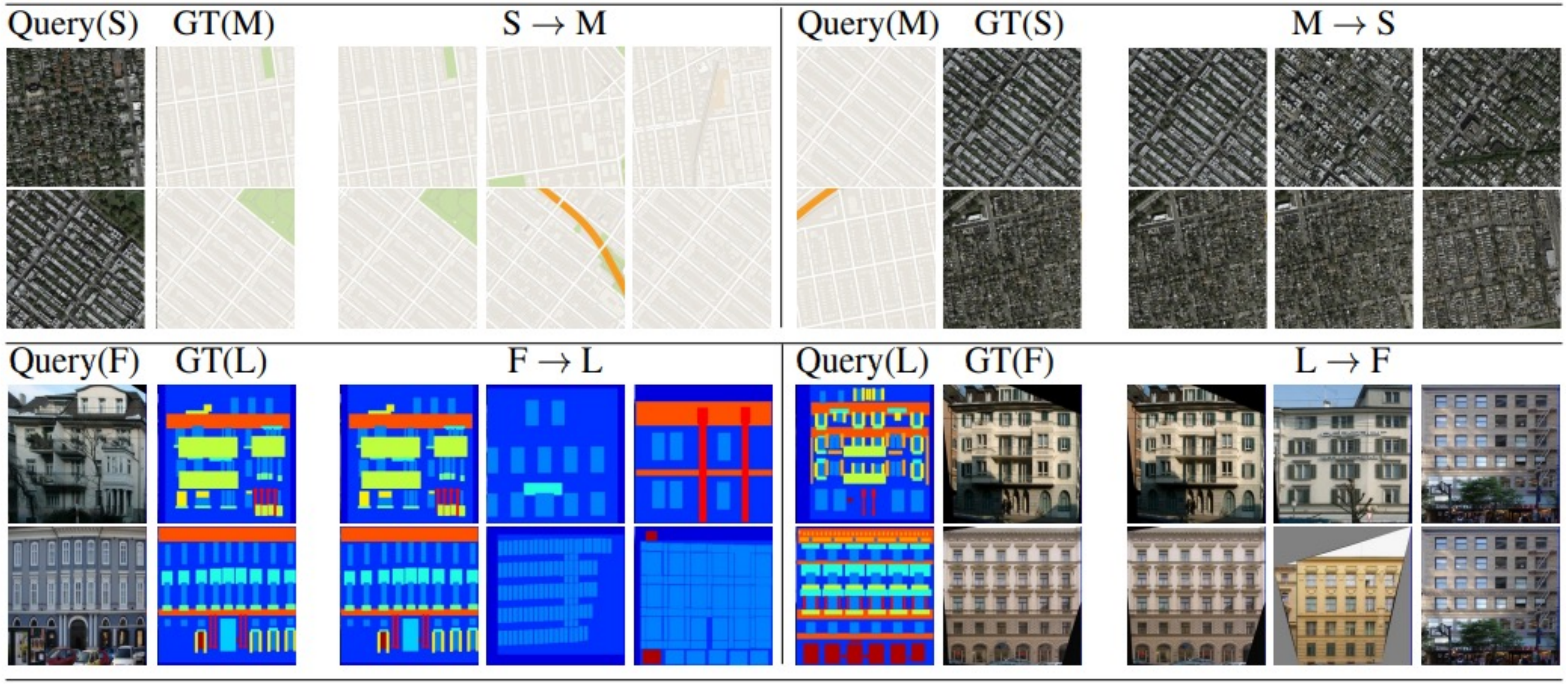
- Cross domain retrieval
  - <u>Maps</u>
    - X : map image
    - Y : satellite image

- <u>Facades</u>
  - X : semantic label map
  - Y : photo of the same building

| 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) |

※ query : $X$, database : $Y$

1. $\mu_x^s = r_\psi(z^s|x).mean$
2. $\mu_y^s = r_\psi(z^s|y).mean$
3. $d\left(\mu_x^s, \mu_y^s\right) \rightarrow \ K\ neighbors$



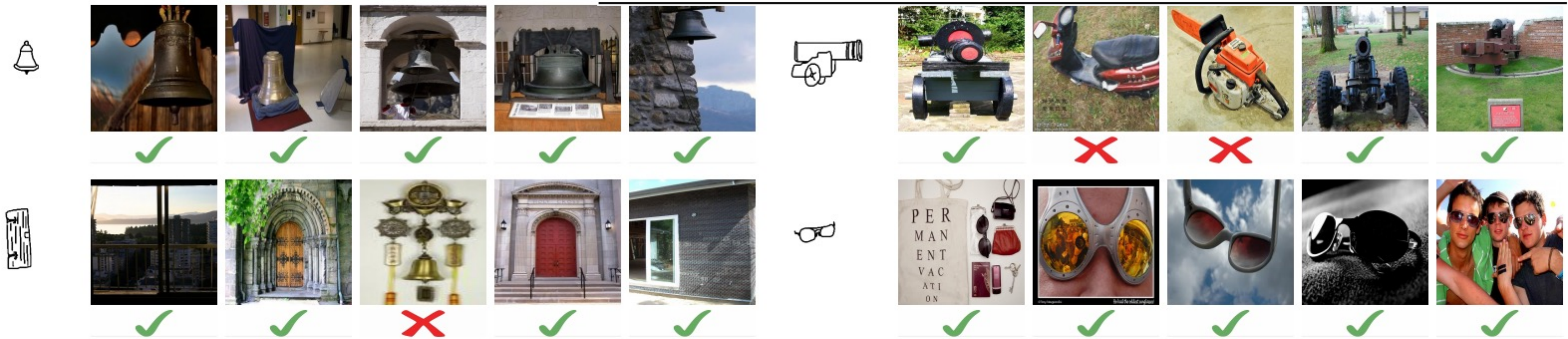
# Experiment

- Zero-shot sketch based image retrieval
  - Sketchy(extended)
    - X : sketches
    - Y : photos

※ query : $X$, database : $Y$

1. $\mu_x^s = r_\psi(z^s|x).mean$
2. $\mu_y^s = r_\psi(z^s|y).mean$
3. $d(\mu_x^s, \mu_y^s) \rightarrow K\ neighbors$

| | Feature | **Evaluation metric** | | **External knowledge** | | |
|---|---|---|---|---|---|---|
| **Models** | Dimension | mAP | P@100 | Attribute | WordEmb. | 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** | - | - | - |



# Summary

- Goal
  - Cross domain disentanglement
  - Partitioning into domain-invariant($z^s$) and domain specific($z^x, z^y$)

- Main idea
  - Further regularization on ELBO relying on the information theory
  - $I(x, y, z^s) - I(z^x, z^s) + I(x, y, z^s) - I(z^y, z^s)$
  - Derive tractable lower bound