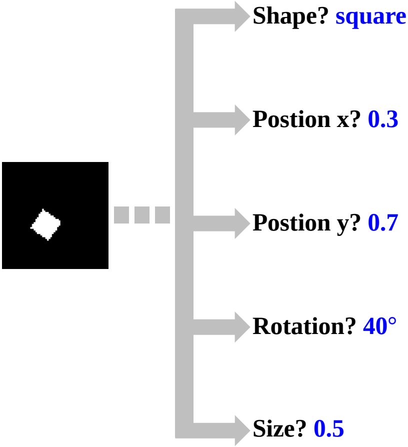
Learning Discrete and Continuous Factors of Data via Alternating Disentanglement

Yeonwoo Jeong, Hyun Oh Song

Seoul National University

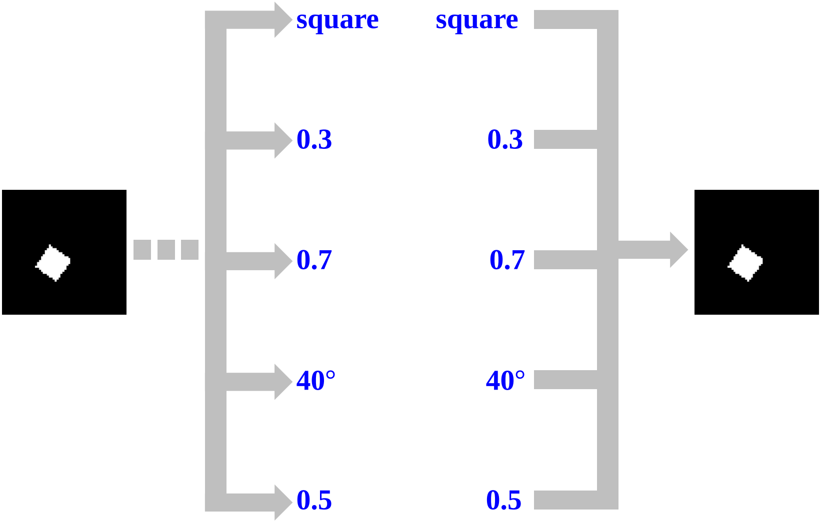
ICML19

Motivation

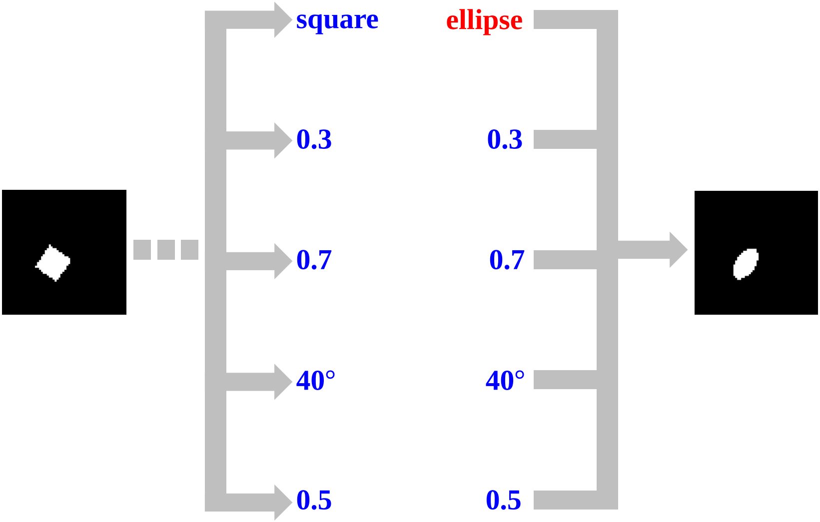


- Our goal is to disentangle the underlying explanatory factors of data without any supervision.

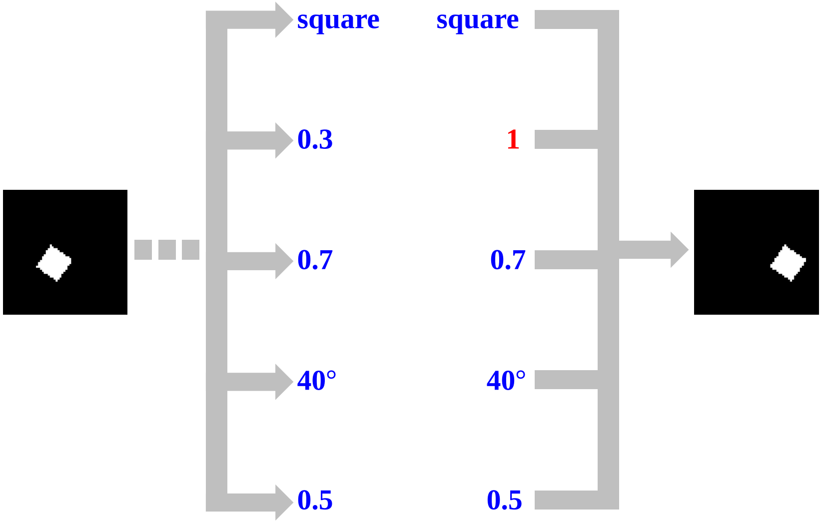
Motivation



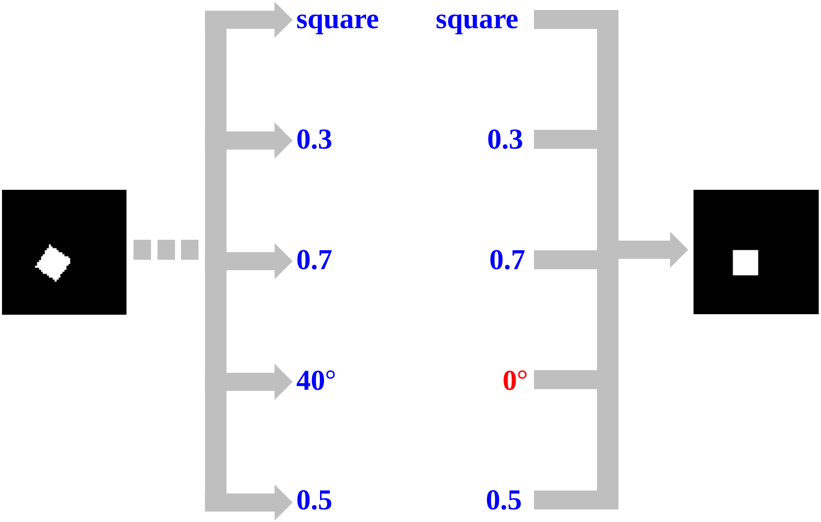
Motivation



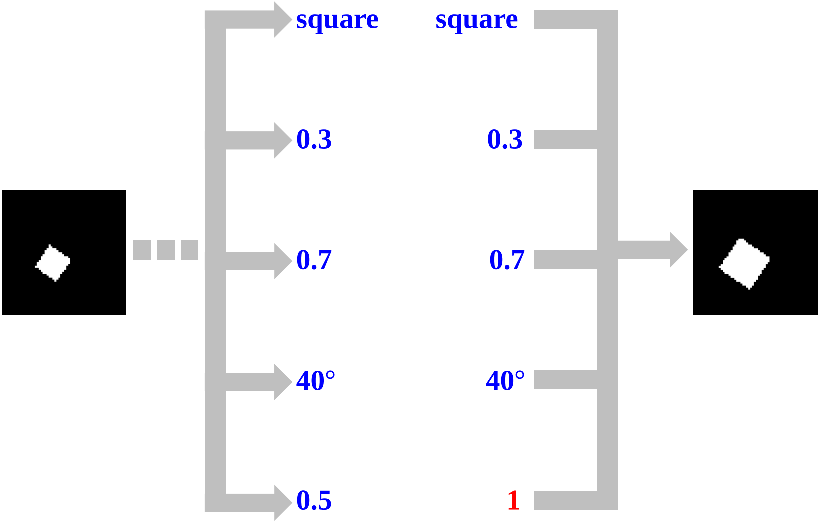
Motivation



Motivation



Motivation



Motivation

- Most recent methods focus on learning only the continuous factors of variation.

Motivation

- Most recent methods focus on learning only the continuous factors of variation.
- **Learning discrete representations** is known as a challenging problem. However, **learning continuous and discrete representations** is a more challenging problem.

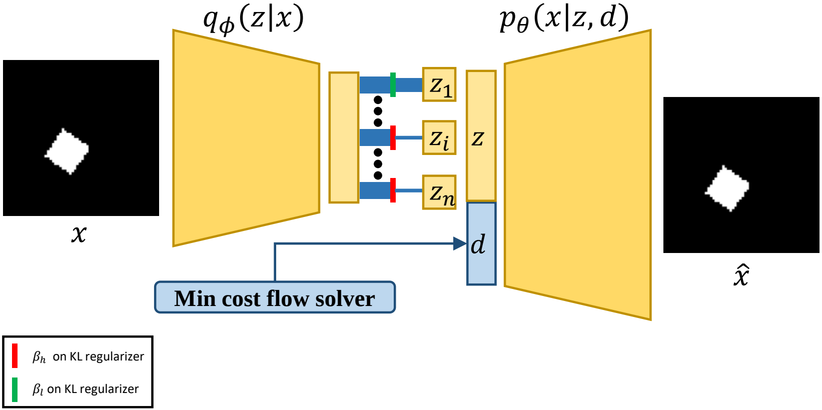
Outline

Method

Experiments

Conclusion

Overview of our method



Overview of our method

- We propose an efficient procedure for implicitly penalizing the total correlation by **controlling the information flow on each variables**.
- We propose a method for jointly learning discrete and continuous latent variables in an **alternating maximization framework**.

Limitation of β-VAE framework

- β-VAE sets $\beta > 1$ to penalize $TC(z)$ for **disentangled representations**.
- However, it penalizes the mutual information$(= I(x, z))$ between the data and the latent variables.

Our method

- We aim at penalizing $TC(z)$ by sequentially penalizing the individual summand $\mathbf{I}(\mathbf{z_{1:i-1}};\mathbf{z_i})$.

$$TC(z) = \sum_{i=2}^{m} \mathbf{I}(\mathbf{z_{1:i-1}};\mathbf{z_i}).$$

Our method

- We aim at penalizing $TC(z)$ by sequentially penalizing the individual summand $\mathbf{I(z_{1:i-1}; z_i)}$.

$$TC(z) = \sum_{i=2}^{m} \mathbf{I(z_{1:i-1}; z_i)}.$$

- We implicitly minimizes each summand, $\mathbf{I(z_{1:i-1}; z_i)}$ by sequentially maximizing the left hand side $I(x; z_{1:i})$ for all $i = 2, \ldots, m$

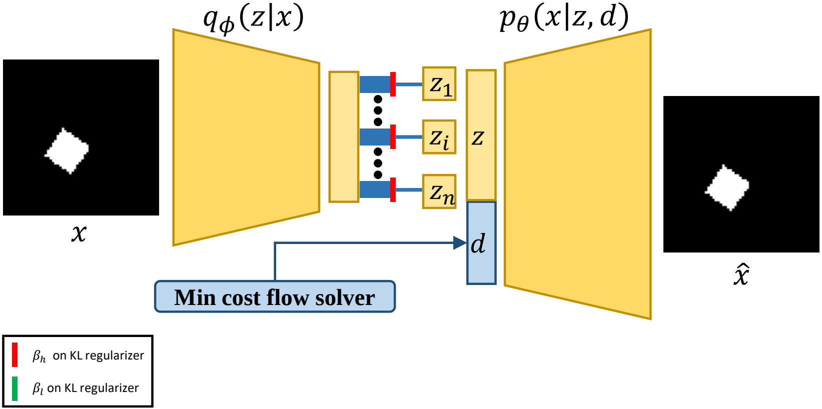
  1.
$$I(x; z_{1:i}) = \underset{\uparrow}{I(x; z_{1:i-1})} + I(x; z_i) - \mathbf{I(z_{1:i-1}; z_i)}.$$

  2.
$$\underset{\uparrow}{I(x; z_{1:i})} = \underset{\bullet}{I(x; z_{1:i-1})} + \underset{\uparrow}{I(x; z_i)} - \underset{\downarrow}{\mathbf{I(z_{1:i-1}; z_i)}}.$$

Our method

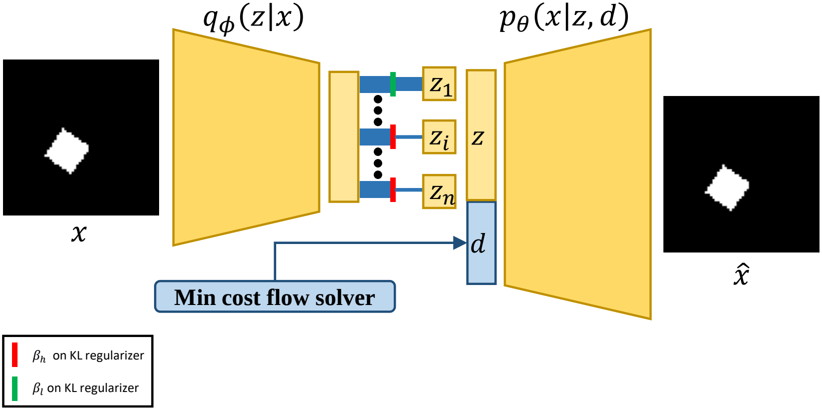
- In practice, we maximize $I(x; z_{1:i})$ by **minimizing reconstruction term** while penalizing $z_{i+1:m}$ with high β $(:= \beta_h)$ and the others with small β $(:= \beta_l)$.

Our method



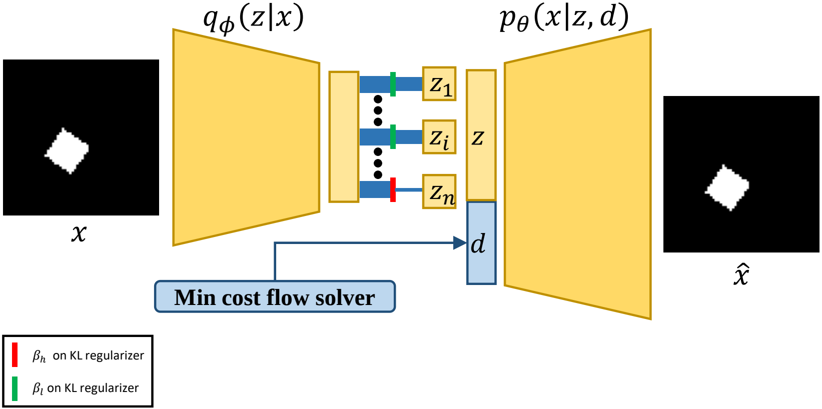
- Every latent dimensions are heavily penalized with β_h. Each penalty on latent dimension is sequentially relieved one at a time with β_l in a **cascading fashion**.

Our method



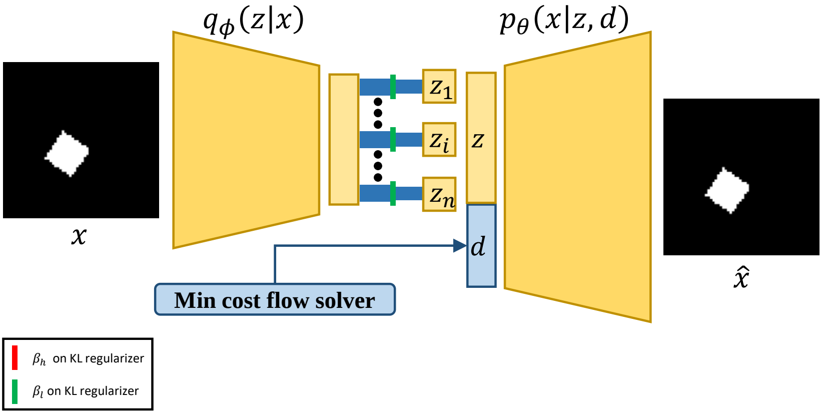
- Every latent dimensions are heavily penalized with β_h. Each penalty on latent dimension is sequentially relieved one at a time with β_l in a **cascading fashion**.

Our method



- Every latent dimensions are heavily penalized with β_h. Each penalty on latent dimension is sequentially relieved one at a time with β_l in a **cascading fashion**.

Our method



- Every latent dimensions are heavily penalized with β_h. Each penalty on latent dimension is sequentially relieved one at a time with β_l in a **cascading fashion**.

Graphical model

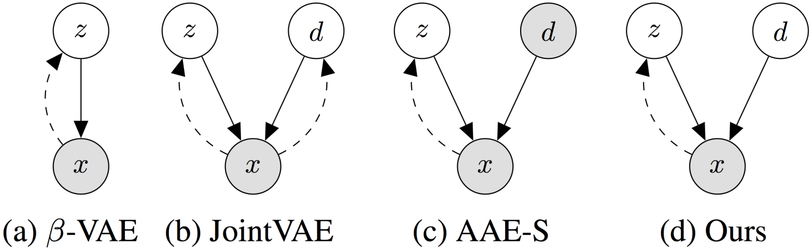


Figure: Graphical models view. **Solid lines** denote the **generative process** and the **dashed lines** denote the **inference process**. x, z, d denotes the data, continuous latent code, and the discrete latent code respectively.

Motviation of our method

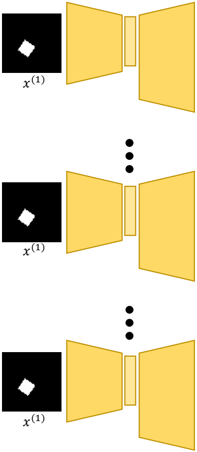
- **AAE with supervised discrete variables(AAE-S)** can learn good continuous representations when the burden of simultaneously modeling the continuous and discrete factors is relieved through supervision on discrete factors unlike **jointVAE**.

Motviation of our method

- **AAE with supervised discrete variables(AAE-S)** can learn good continuous representations when the burden of simultaneously modeling the continuous and discrete factors is relieved through supervision on discrete factors unlike **jointVAE**.

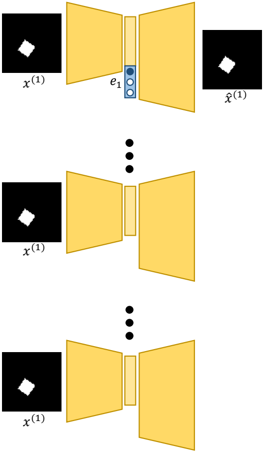
- Inspired by these findings, our idea is to **alternate** between finding the most likely discrete configuration of the variables given the continuous factors, and updating the parameters (ϕ, θ) given the discrete configurations.

Construct unary term



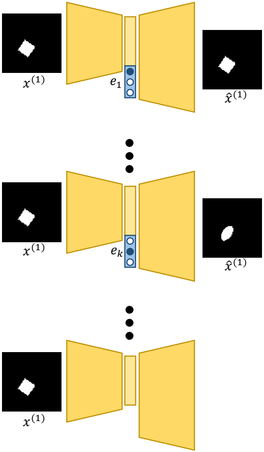
- The discrete latent variables are represented using one-hot encodings of each variables $d^{(i)} \in \{e_1, \ldots, e_S\}$.

Construct unary term



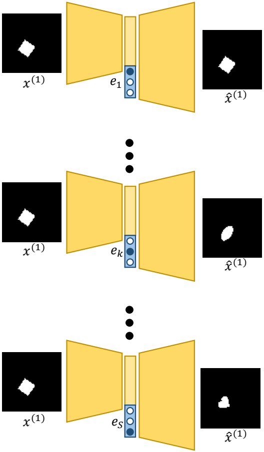
- The discrete latent variables are represented using one-hot encodings of each variables $d^{(i)} \in \{e_1, \ldots, e_S\}$.

Construct unary term



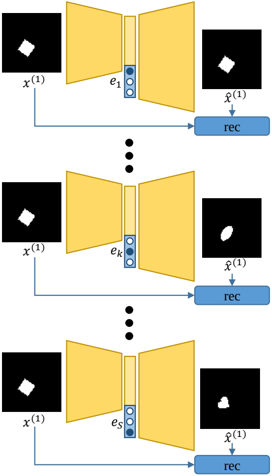
- The discrete latent variables are represented using one-hot encodings of each variables $d^{(i)} \in \{e_1, \ldots, e_S\}$.

Construct unary term



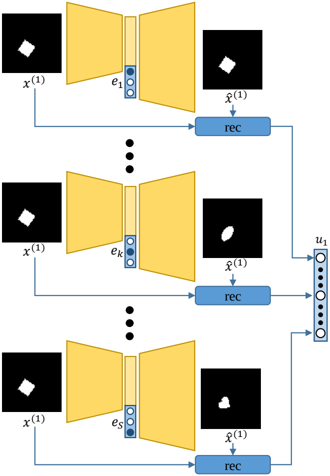
- The discrete latent variables are represented using one-hot encodings of each variables $d^{(i)} \in \{e_1, \ldots, e_S\}$.

Construct unary term



- The discrete latent variables are represented using one-hot encodings of each variables $d^{(i)} \in \{e_1, \ldots, e_S\}$.

Construct unary term



- The discrete latent variables are represented using one-hot encodings of each variables $d^{(i)} \in \{e_1, \ldots, e_S\}$.
- $u_\theta^{(i)}$ denotes the vector of the likelihood $\log p_\theta(x^{(i)}|z^{(i)}, e_k)$ evaluated at each $k \in [S]$.

Alternating minimization scheme

- Our goal is to maximize the variational lower bound of the following objective,

$$\mathcal{L}(\theta, \phi) = I(x; [z, d]) - \beta \mathbb{E}_{x \sim p(x)} D_{\mathsf{KL}}(q_\phi(z \mid x) \parallel p(z)) - \lambda D_{\mathsf{KL}}(q(d) \parallel p(d))$$

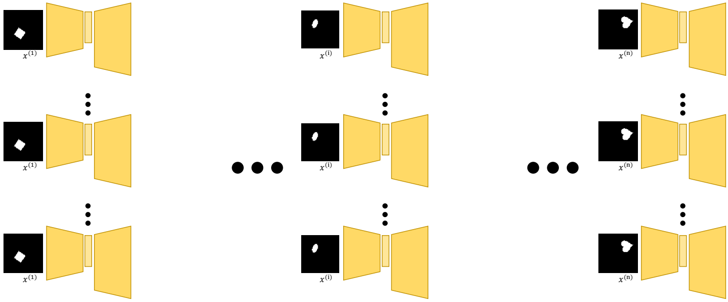
- After rearranging the terms, we arrive at the following optimization problem.

$$\underset{\theta,\phi}{\text{maximize}} \left(\underbrace{\underset{d^{(1)},\ldots d^{(n)}}{\text{maximize}} \sum_{i=1}^{n} {u_\theta^{(i)}}^{\mathsf{T}} d^{(i)} - \lambda' \sum_{i \neq j} {d^{(i)}}^{\mathsf{T}} d^{(j)}}_{:=\mathcal{L}_{LB}(\theta,\phi)} \right)$$

$$- \beta \sum_{i=1}^{n} D_{KL}(q_\phi(z|x^{(i)})||p(z))$$

$$\text{subject to} \quad \|d^{(i)}\|_1 = 1, \; d^{(i)} \in \{0,1\}^S, \; \forall i,$$

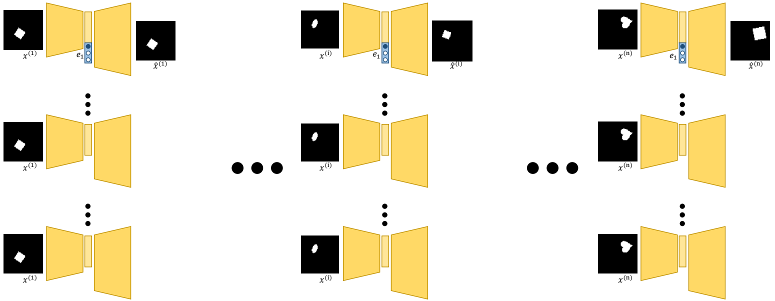
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \ldots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

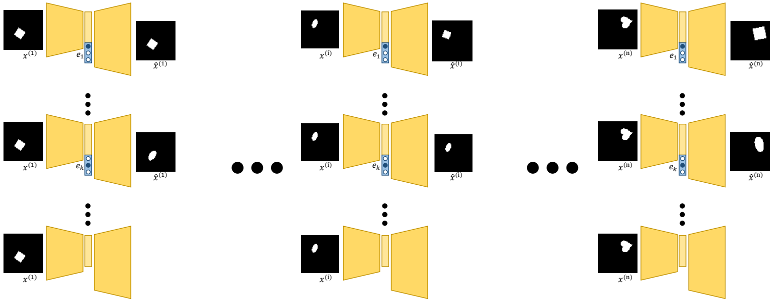
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \ldots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

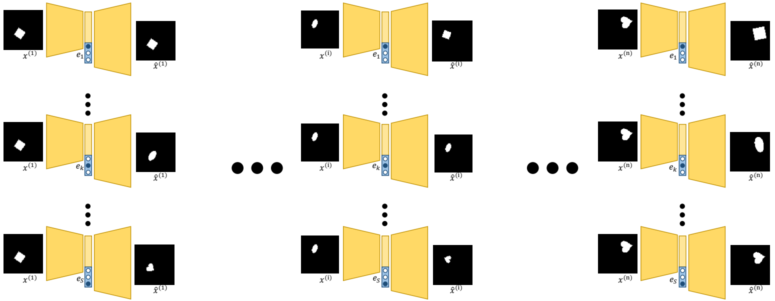
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \ldots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

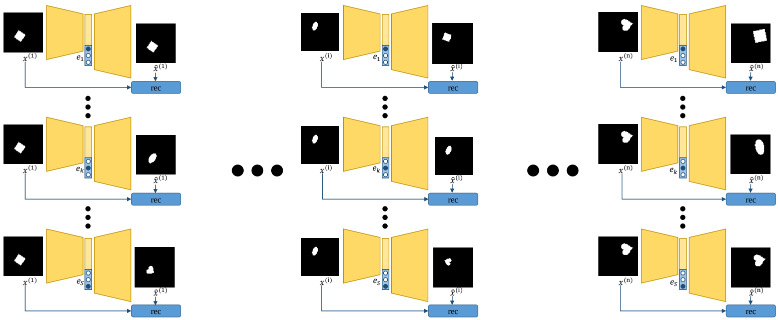
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \dots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

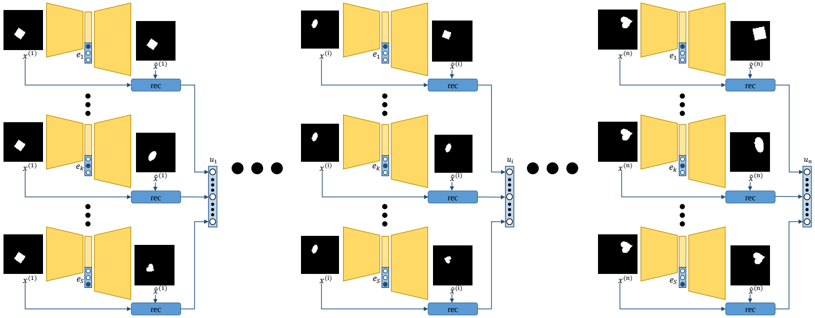
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \ldots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

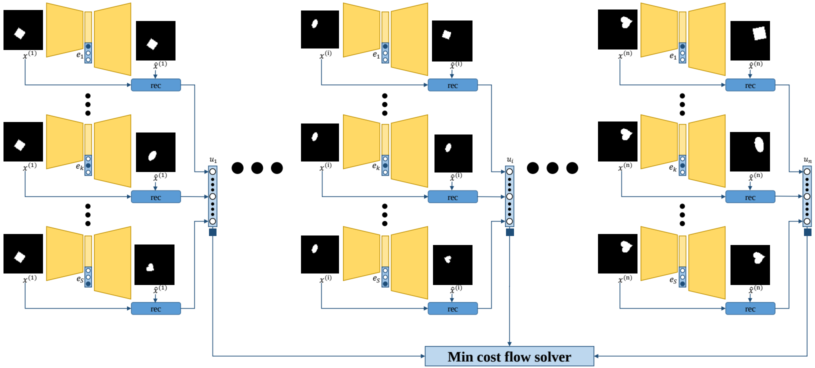
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \ldots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

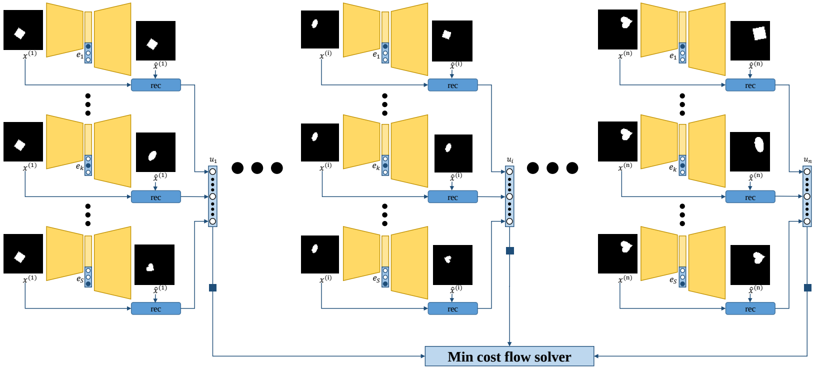
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \ldots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

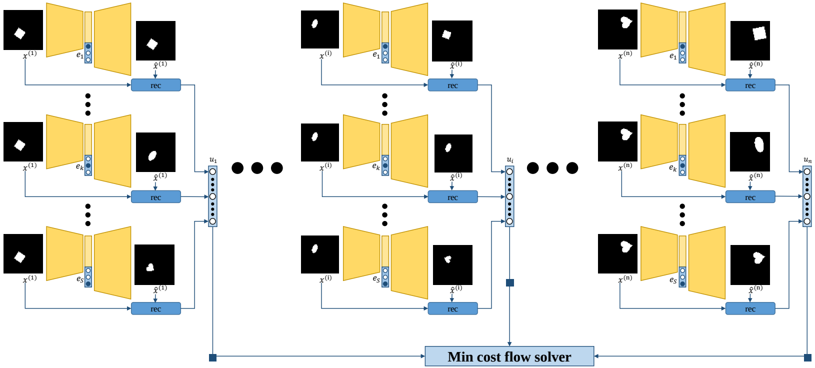
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \ldots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

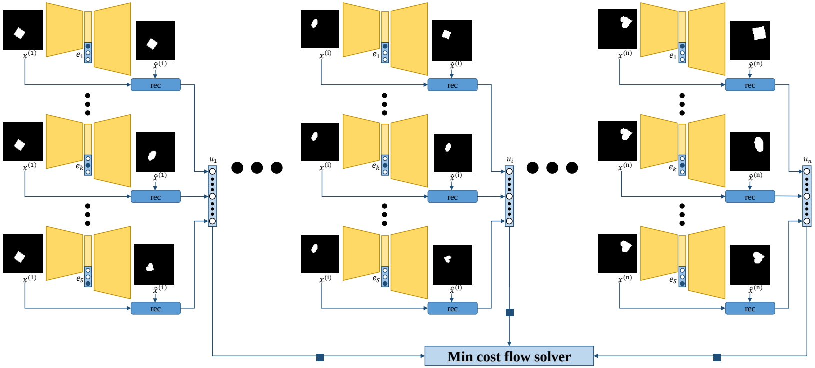
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \ldots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

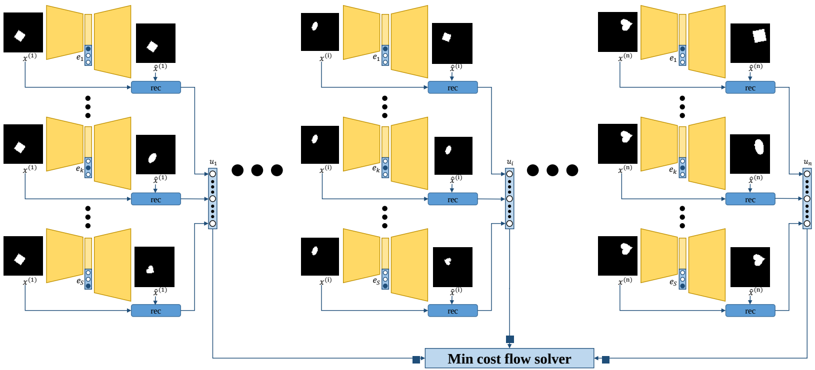
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \ldots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

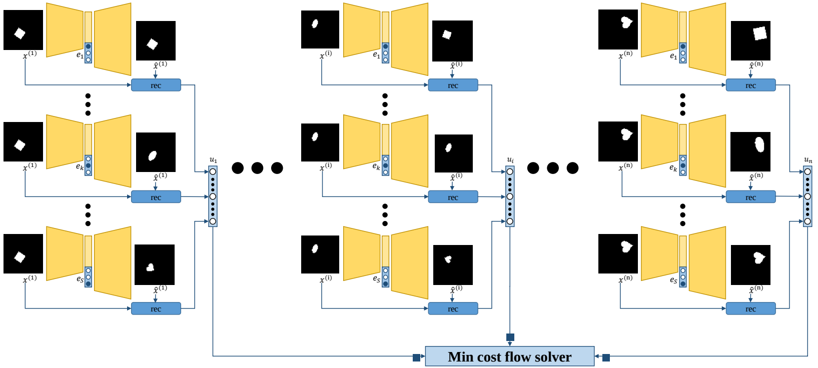
Finding the most likely discrete configuration



- With the unary terms, we solve inner maximization problem $\mathcal{L}_{LB}(\theta, \phi)$ over the discrete variables $[d^{(1)}, \ldots, d^{(n)}]$.[1]

[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

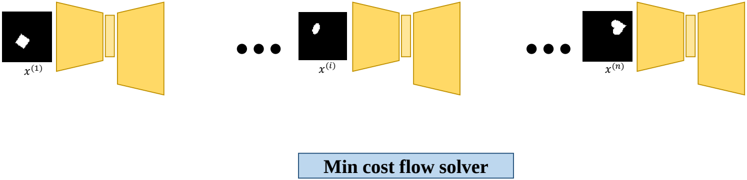
Finding the most likely discrete configuration



- The maximization problem can be exactly solved in polynomial time via *minimum cost flow*(mcf) without continuous relaxation.[1]

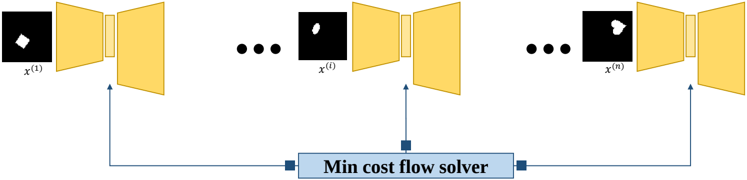
[1] Jeong, Y. and Song, H. O. "Efficient end-to-end learning for quantizable representations" ICML2018.

Updating the parameters



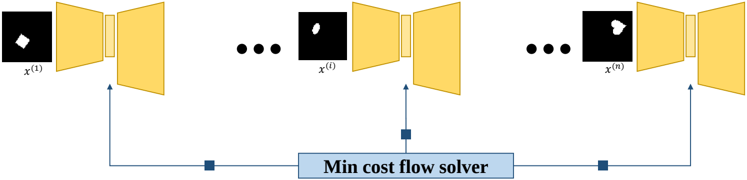
- Then, we update the parameters under this discrete configurations.

Updating the parameters



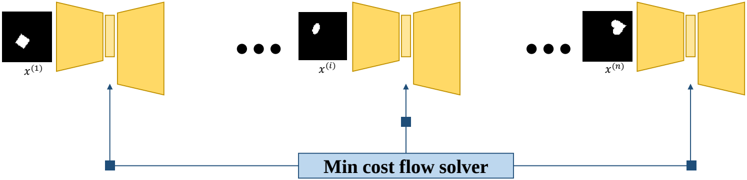
- Then, we update the parameters under this discrete configurations.

Updating the parameters



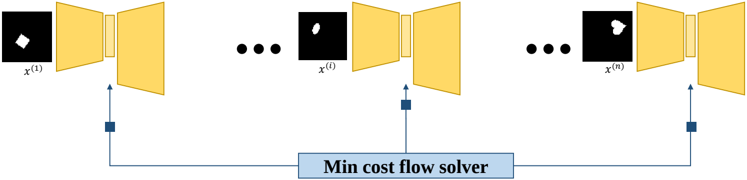
- Then, we update the parameters under this discrete configurations.

Updating the parameters



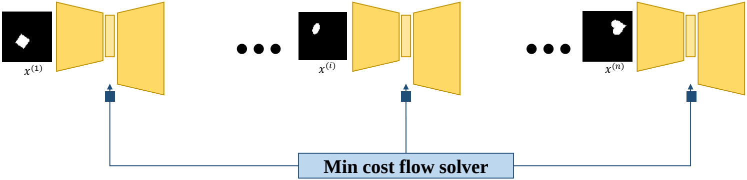
- Then, we update the parameters under this discrete configurations.

Updating the parameters



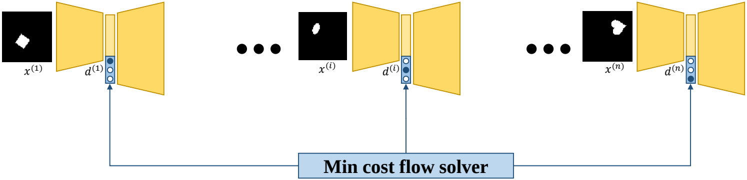
- Then, we update the parameters under this discrete configurations.

Updating the parameters



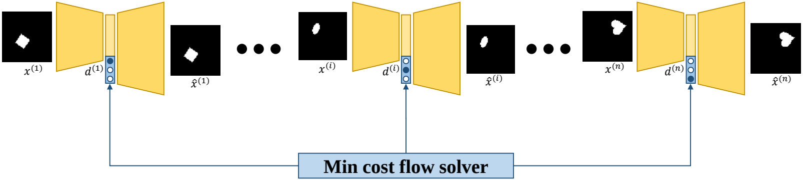
- Then, we update the parameters under this discrete configurations.

Updating the parameters



- Then, we update the parameters under this discrete configurations.

Updating the parameters



- Then, we update the parameters under this discrete configurations.

Outline

Method

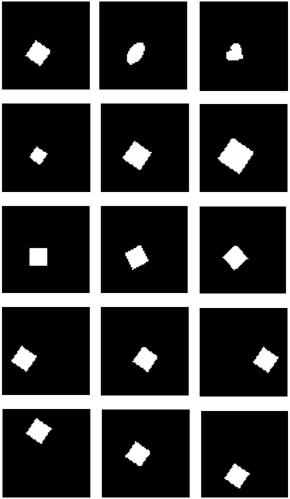
Experiments

Conclusion

Notation

- We denote our full method as **CascadeVAE**.
- We evaluate with disentanglement score introduced in **FactorVAE** and unsupervised classification accuracy.
- Baselines are β-**VAE, JointVAE, FactorVAE**

dSprites Dataset Example



- Shape (discrete) : square, ellipse, heart
- Scale: 6 values linearly spaced in $[0.5, 1]$
- Orientation: 40 values in $[0, 2\pi]$
- Position X: 32 values in $[0, 1]$
- Position Y: 32 values in $[0, 1]$

Quantitative results on dSprites

Disentanglement score

Method	m	Mean (std)	Best
β VAE			
$(\beta = 10.0)$	5	70.11 (7.54)	84.62
$(\beta = 4.0)$	10	74.41 (7.68)	88.38
FactorVAE	5	81.09 (2.63)	85.12
	10	82.15 (0.88)	88.25
JointVAE	6	74.51 (5.17)	91.75
	4	73.06 (2.18)	75.38
CascadeVAE			
$(\beta_l = 1.0)$	6	90.49 (5.28)	**99.50**
$(\beta_l = 2.0)$	4	**91.34 (7.36)**	98.62

Unsupervised classification accuracy

Method	m	Mean (std)	Best
JointVAE	6	44.79 (3.88)	53.14
	4	43.99 (3.94)	54.11
CascadeVAE	6	**78.84 (15.65)**	**99.66**
	4	76.00 (22.16)	98.72

Outline

Method

Experiments

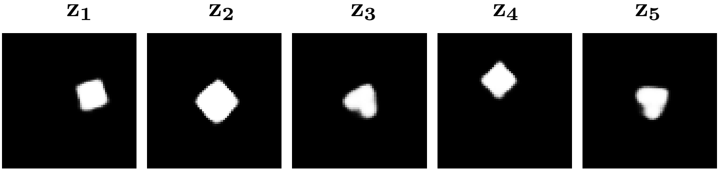
Conclusion

Conclusion

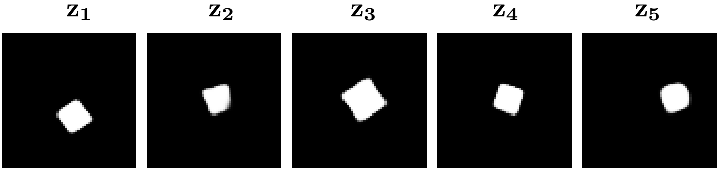
- Our experiments show that information cascading and alternating maximization of discrete and continuous variables, lead to the state of the art performance in 1) **disentanglement score**, and 2) **classification accuracy**.
- The source code is available at https://github.com/snu-mllab/DisentanglementICML19.

Latent dimension traversal in dSprites

β-VAE

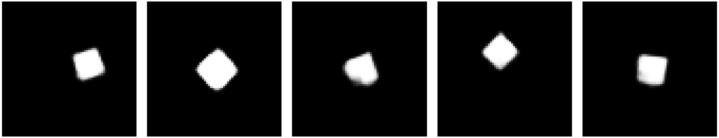


FactorVAE



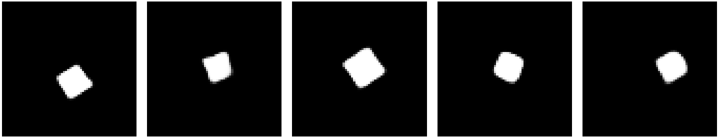
β-VAE

$\mathbf{z_1}$ $\mathbf{z_2}$ $\mathbf{z_3}$ $\mathbf{z_4}$ $\mathbf{z_5}$

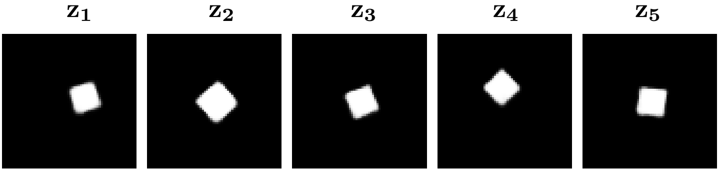


FactorVAE

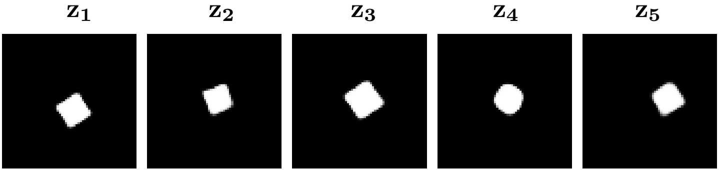
$\mathbf{z_1}$ $\mathbf{z_2}$ $\mathbf{z_3}$ $\mathbf{z_4}$ $\mathbf{z_5}$



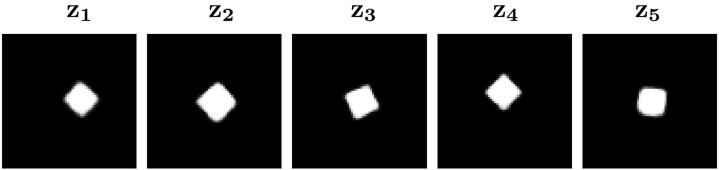
β-VAE



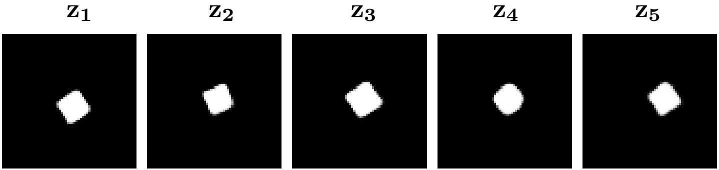
FactorVAE



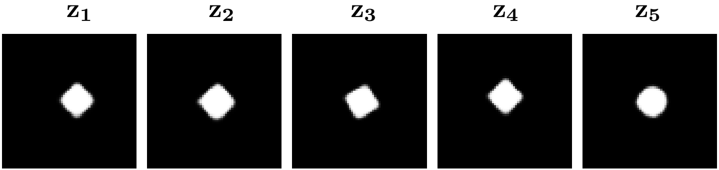
β-VAE



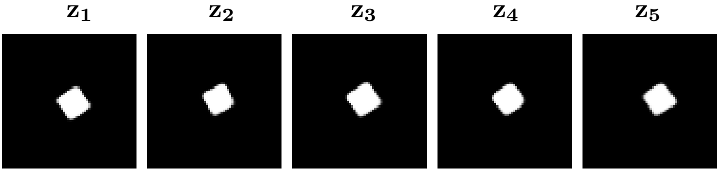
FactorVAE



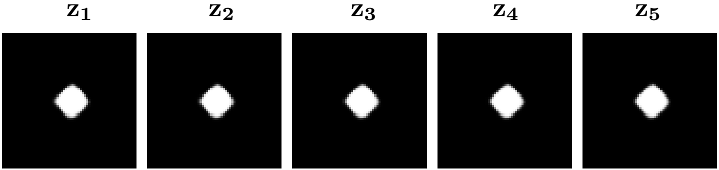
β-VAE



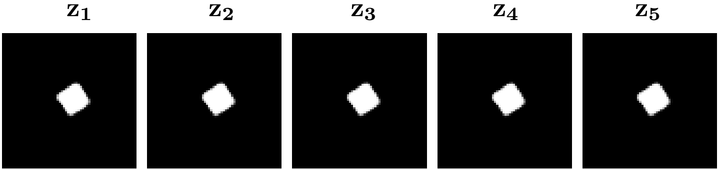
FactorVAE



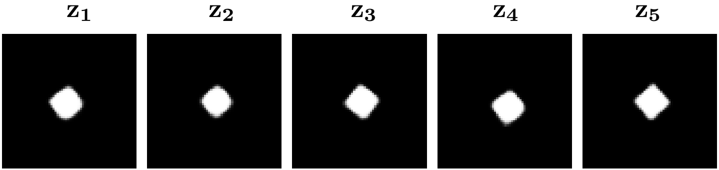
β-VAE



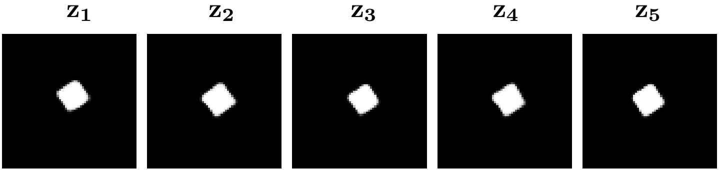
FactorVAE



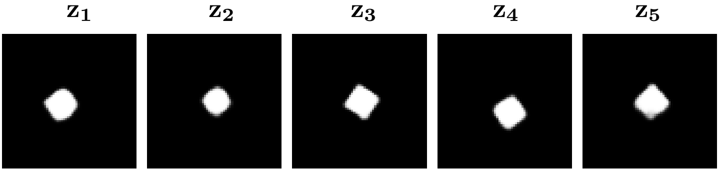
β-VAE



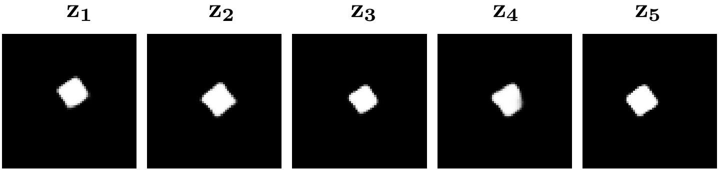
FactorVAE



β-VAE

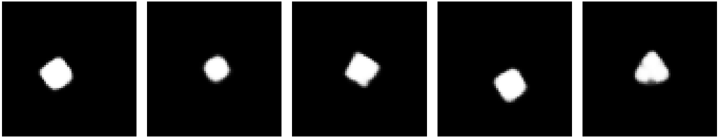


FactorVAE



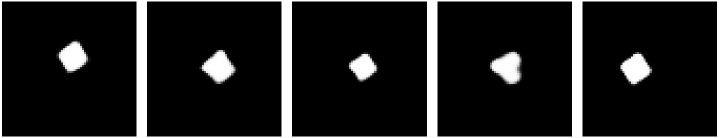
β-VAE

$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$

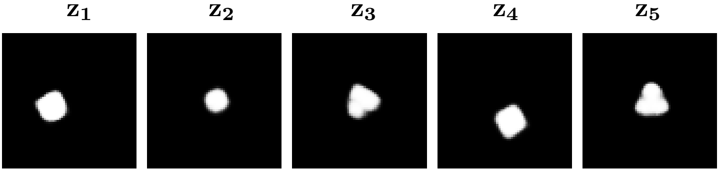


FactorVAE

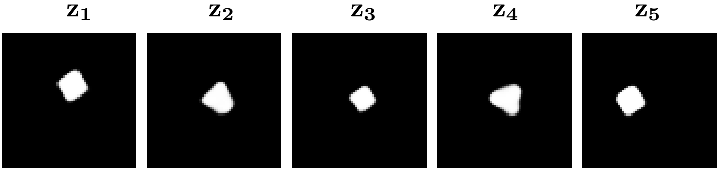
$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$



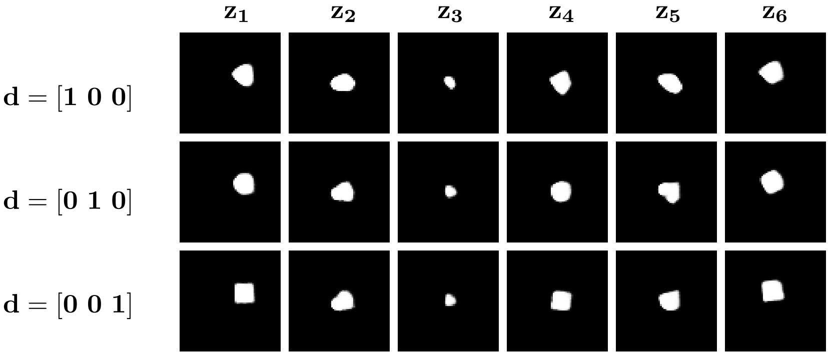
β-VAE



FactorVAE

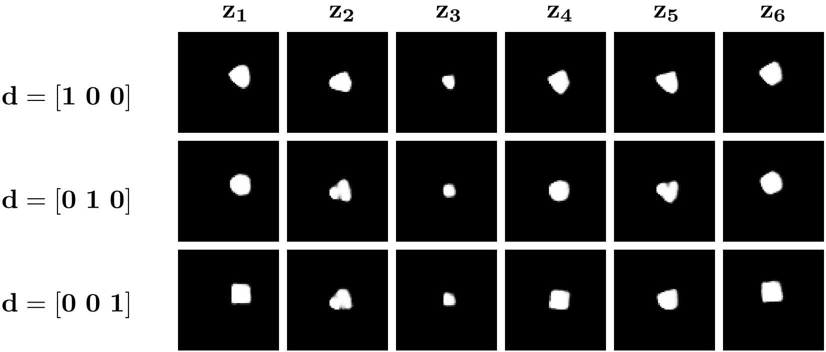


JointVAE



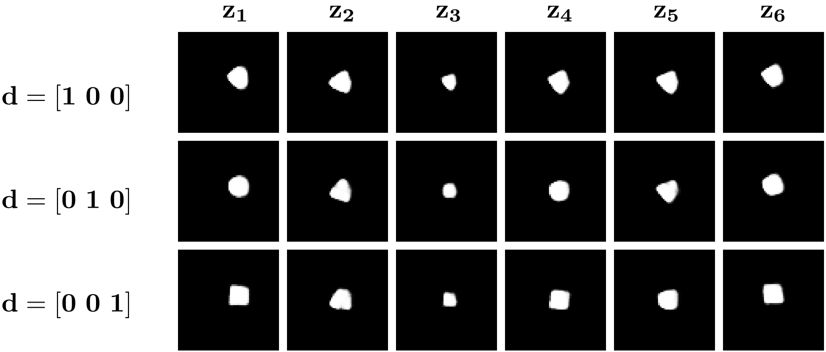
$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

JointVAE



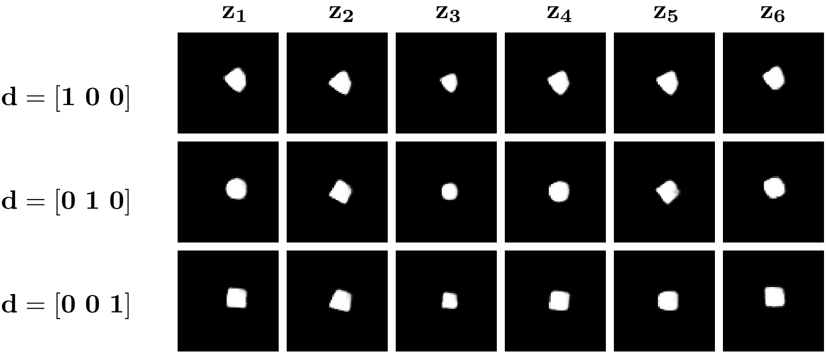
$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

JointVAE



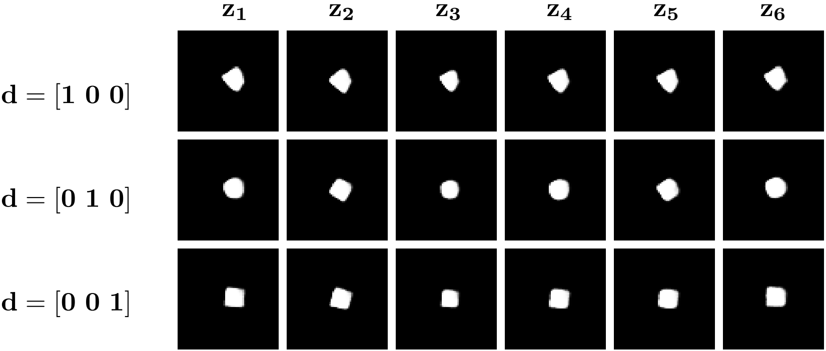
$\mathbf{z_1}$ $\mathbf{z_2}$ $\mathbf{z_3}$ $\mathbf{z_4}$ $\mathbf{z_5}$ $\mathbf{z_6}$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

JointVAE



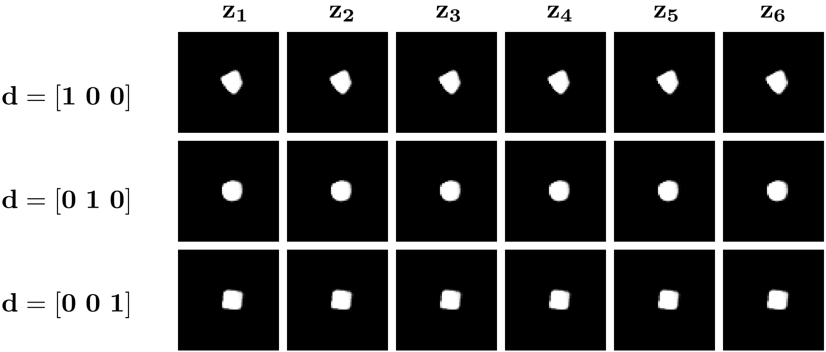
$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [\mathbf{1}\ \mathbf{0}\ \mathbf{0}]$
$\mathbf{d} = [\mathbf{0}\ \mathbf{1}\ \mathbf{0}]$
$\mathbf{d} = [\mathbf{0}\ \mathbf{0}\ \mathbf{1}]$

JointVAE

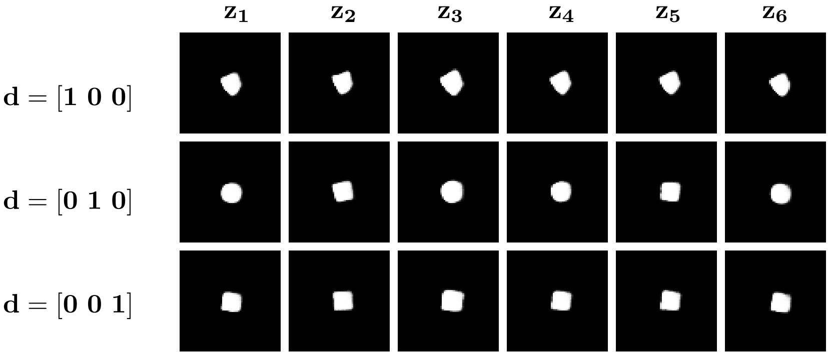


$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

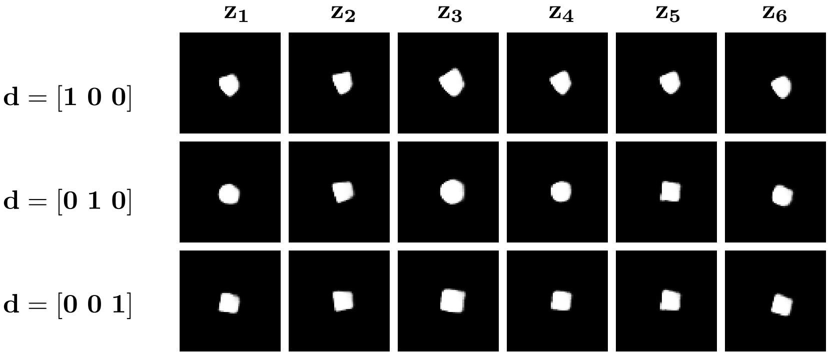
JointVAE


$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$ $\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

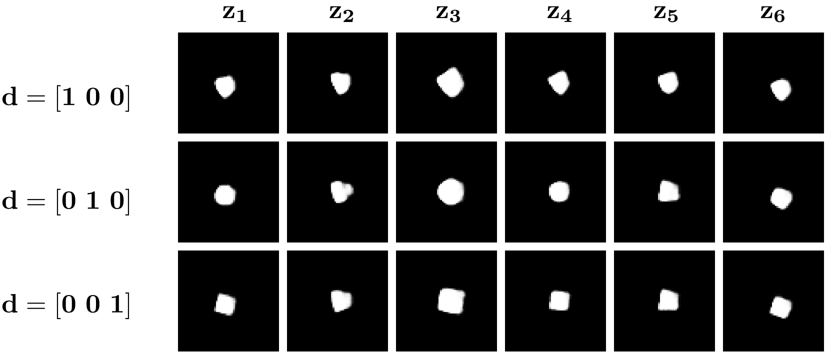
JointVAE


$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

JointVAE

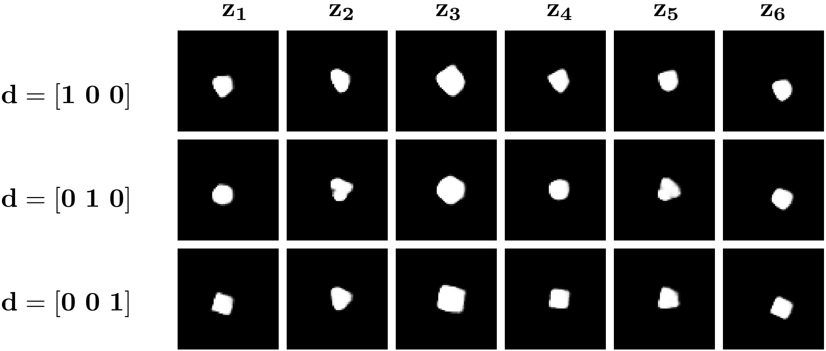

$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$ $\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

JointVAE

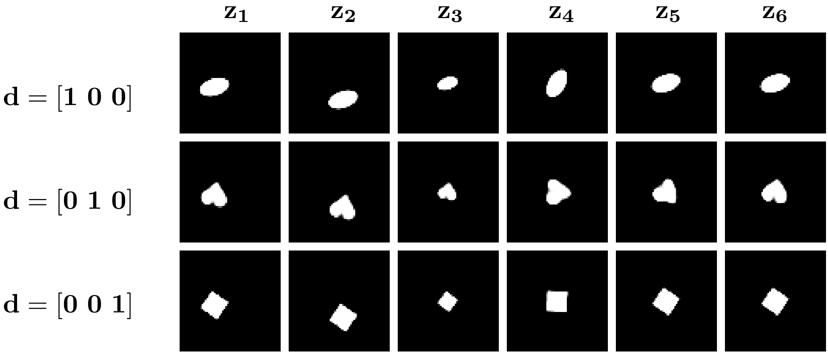


$\mathbf{z_1}$ $\mathbf{z_2}$ $\mathbf{z_3}$ $\mathbf{z_4}$ $\mathbf{z_5}$ $\mathbf{z_6}$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

JointVAE

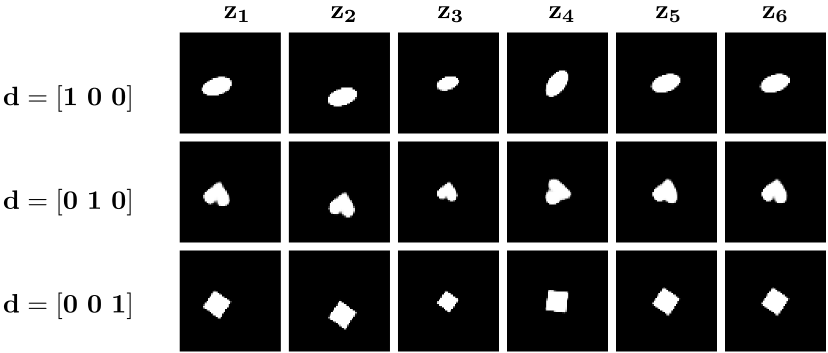

$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

CascadeVAE



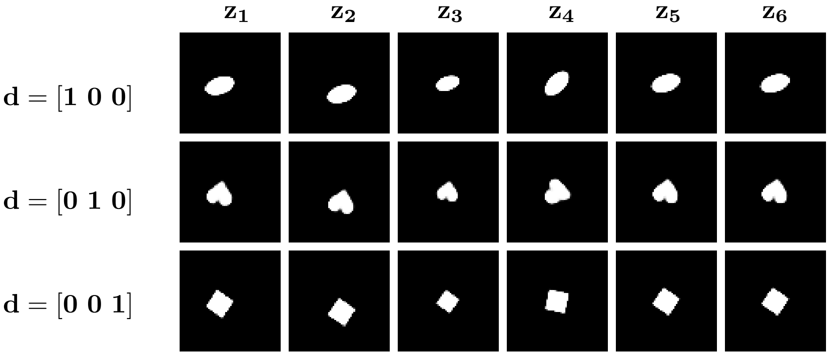
$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$ $\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

CascadeVAE



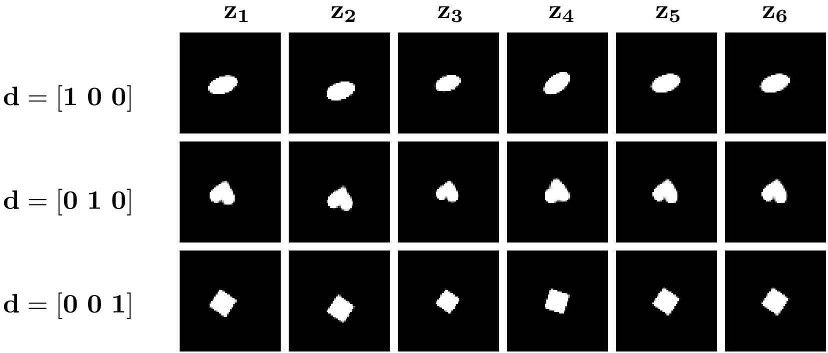
$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$ $\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

CascadeVAE



$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

CascadeVAE


$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

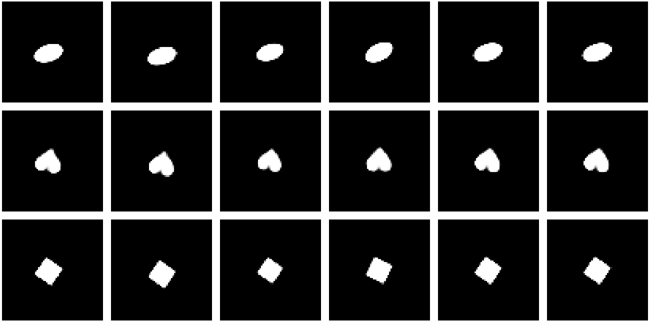
CascadeVAE

$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$ $\mathbf{z}_6$

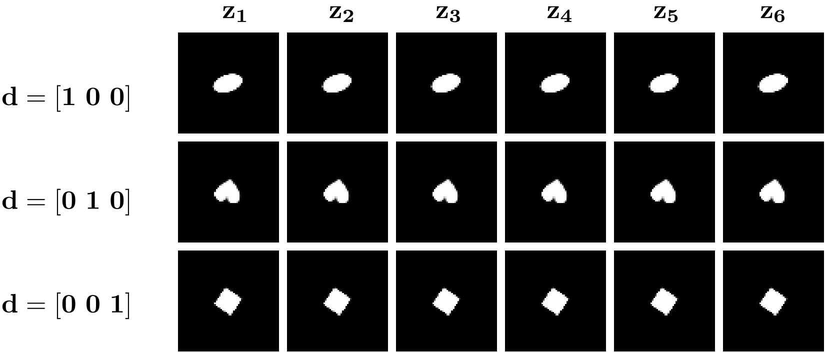
$\mathbf{d} = [1\ 0\ 0]$

$\mathbf{d} = [0\ 1\ 0]$

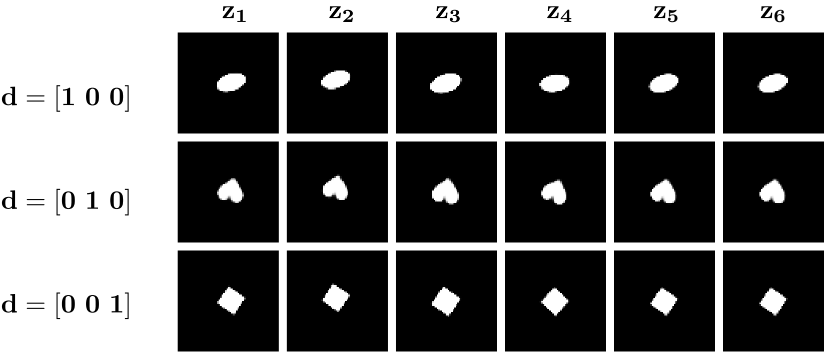
$\mathbf{d} = [0\ 0\ 1]$



CascadeVAE

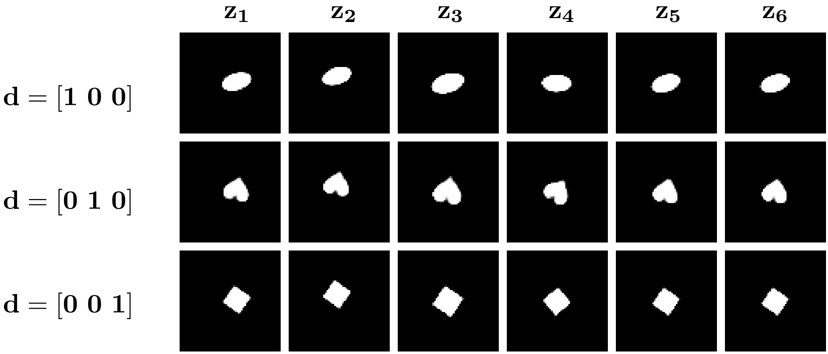

$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$ $\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

CascadeVAE



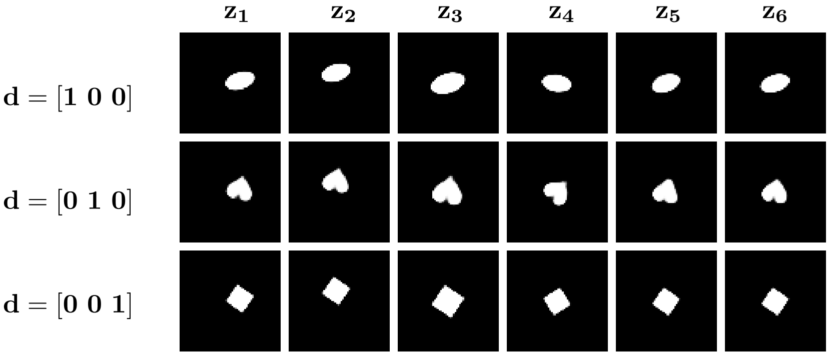
$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

CascadeVAE



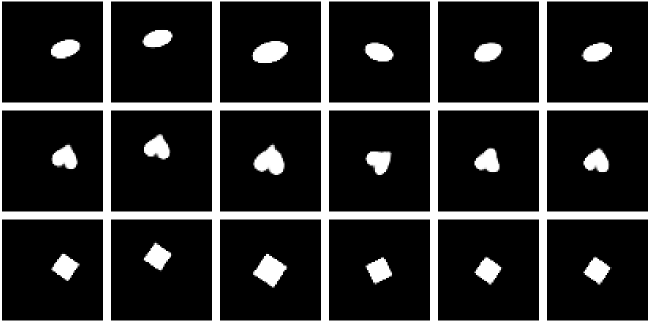
$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

CascadeVAE



$\mathbf{z}_1$
$\mathbf{z}_2$
$\mathbf{z}_3$
$\mathbf{z}_4$
$\mathbf{z}_5$
$\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$

CascadeVAE



$\mathbf{z}_1$ $\mathbf{z}_2$ $\mathbf{z}_3$ $\mathbf{z}_4$ $\mathbf{z}_5$ $\mathbf{z}_6$
$\mathbf{d} = [1\ 0\ 0]$
$\mathbf{d} = [0\ 1\ 0]$
$\mathbf{d} = [0\ 0\ 1]$