Soft-IntroVAE: Analyzing and Improving the Introspective Variational Autoencoder

CVPR 2021 (Oral)

Tal Daniel, Aviv Tamar

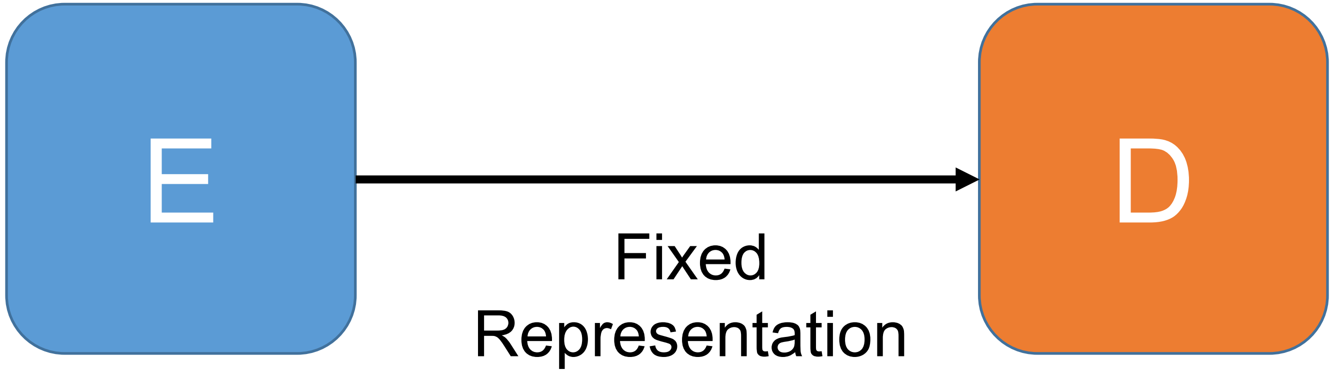
Department of Electrical Engineering Technion, Haifa, Israel

Outline

- Background
- Method
- Experiments
- Conclusion

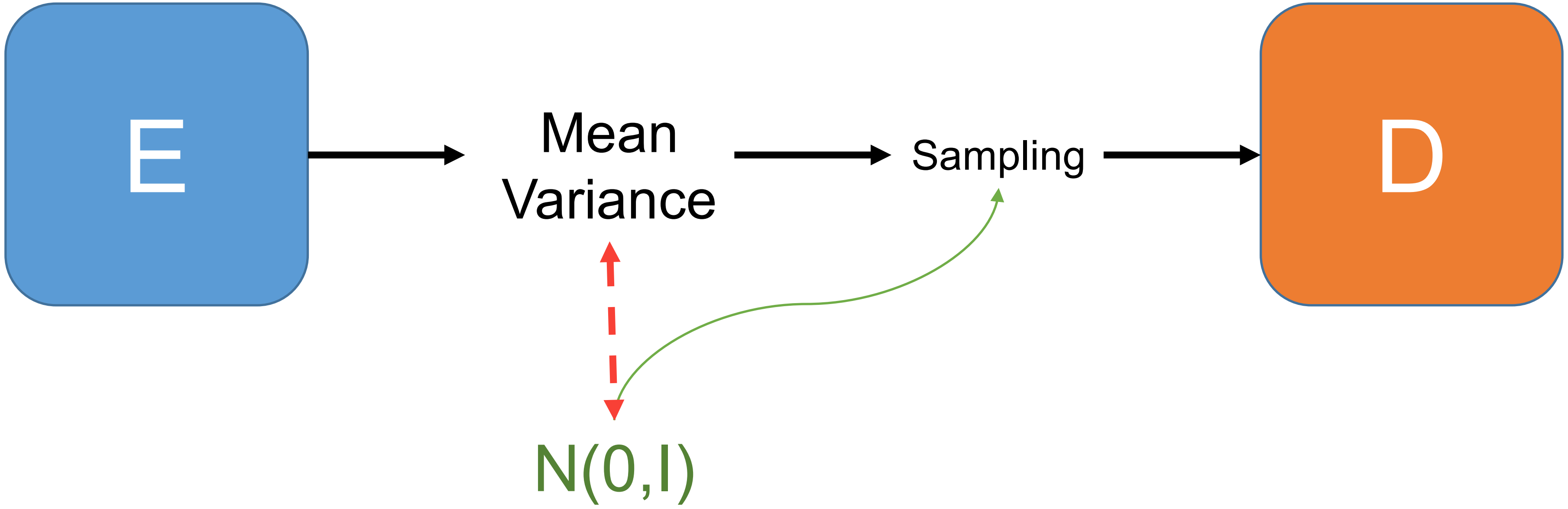
Auto-Encoder

- Compression



Varitional Auto-Encoder (VAE)

- Generation



VAE

- ELBO

Reconstruction

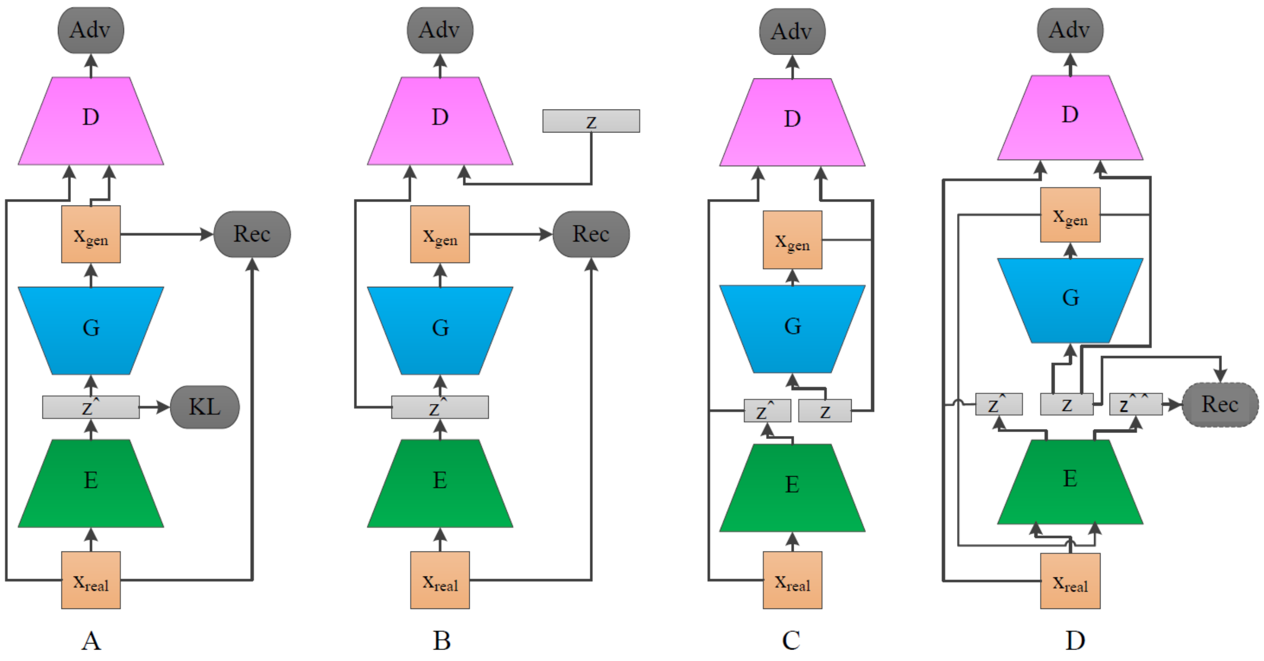
Distribution Constraint

$$\log p_\theta(x) \geq \mathbb{E}_{q(z|x)} \left[\log p_\theta(x|z)\right] - KL(q(z|x)\|p(z)) \\ \doteq ELBO(x), \tag{1}$$

Prior (Gaussian)

Hybrid VAE

- VAE is blurry
  - Learning from GAN



  - Too Complex

IntroVAE

- Real → minimizing KL ; Generated → maximizing KL
  - Encoder: Maximizing KL of generated samples
  - Decoder: Fooling the Encoder → Generating samples minimizing KL
  - Nash equilibrium
    - p_G = p_data

KL between posterior and prior

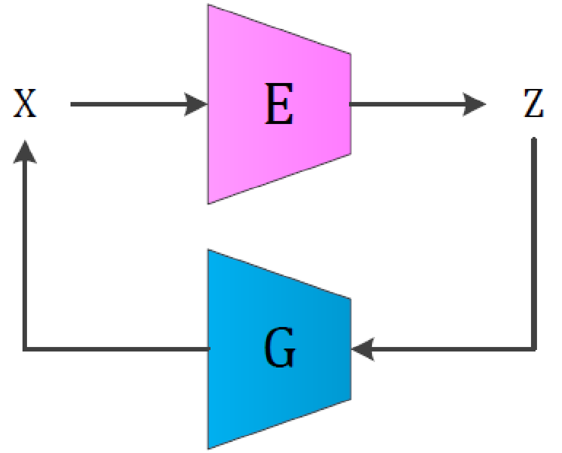
Encoder $$L_E(x, z) = E(x) + [m - E(G(z))]^+ + L_{AE}(x),$$

Hard Threshold

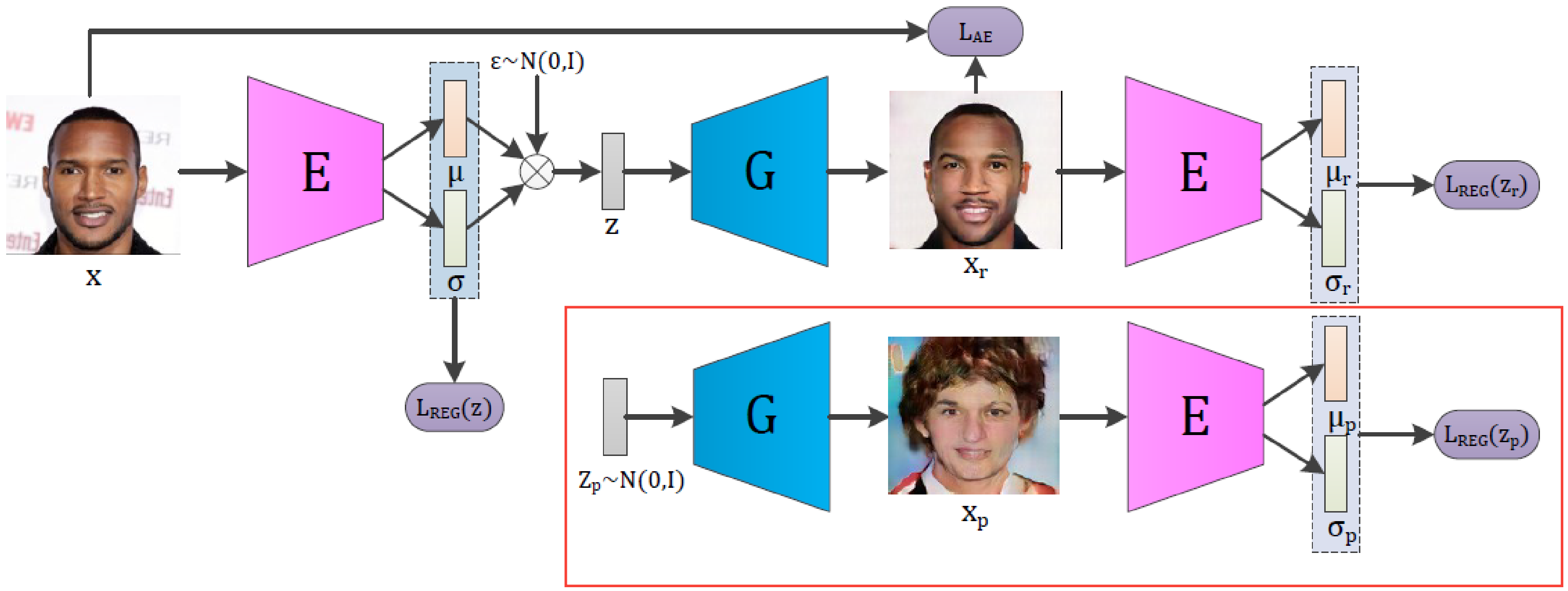
Reconstruction

Decoder $$L_G(z) = E(G(z)) + L_{AE}(x).$$

IntroVAE



X
E
Z
G



L_{AE}
ε~N(0,I)
x
E
μ
σ
z
G
x_r
E
μ_r
σ_r
$L_{REG}(z_r)$
$L_{REG}(z)$
Z_p~N(0,I)
G
x_p
E
μ_p
σ_p
$L_{REG}(z_p)$

IntroVAE

Algorithm 1 Training IntroVAE model

$\theta_G, \phi_E \leftarrow$ Initialize network parameters
while not converged **do**
  $X \leftarrow$ Random mini-batch from dataset
  $Z \leftarrow Enc(X)$
  $Z_p \leftarrow$ Samples from prior $N(0, I)$
  $X_r \leftarrow Dec(Z),\ X_p \leftarrow Dec(Z_p)$
  $L_{AE} \leftarrow L_{AE}(X_r, X)$
  $Z_r \leftarrow Enc(ng(X_r)),\ Z_{pp} \leftarrow Enc(ng(X_p))$
  $L_{adv}^E \leftarrow L_{REG}(Z) + \alpha\{[m - L_{REG}(Z_r)]^+ + [m - L_{REG}(Z_{pp})]^+\}$
  $\phi_E \leftarrow \phi_E - \eta \nabla_{\phi_E}(L_{adv}^E + \beta L_{AE})$ ▷ Perform Adam updates for ϕ_E
  $Z_r \leftarrow Enc(X_r),\ Z_{pp} \leftarrow Enc(X_p)$
  $L_{adv}^G \leftarrow \alpha\{L_{REG}(Z_r) + L_{REG}(Z_{pp})\}$
  $\theta_G \leftarrow \theta_G - \eta \nabla_{\theta_G}(L_{adv}^G + \beta L_{AE})$ ▷ Perform Adam updates for θ_G
end while

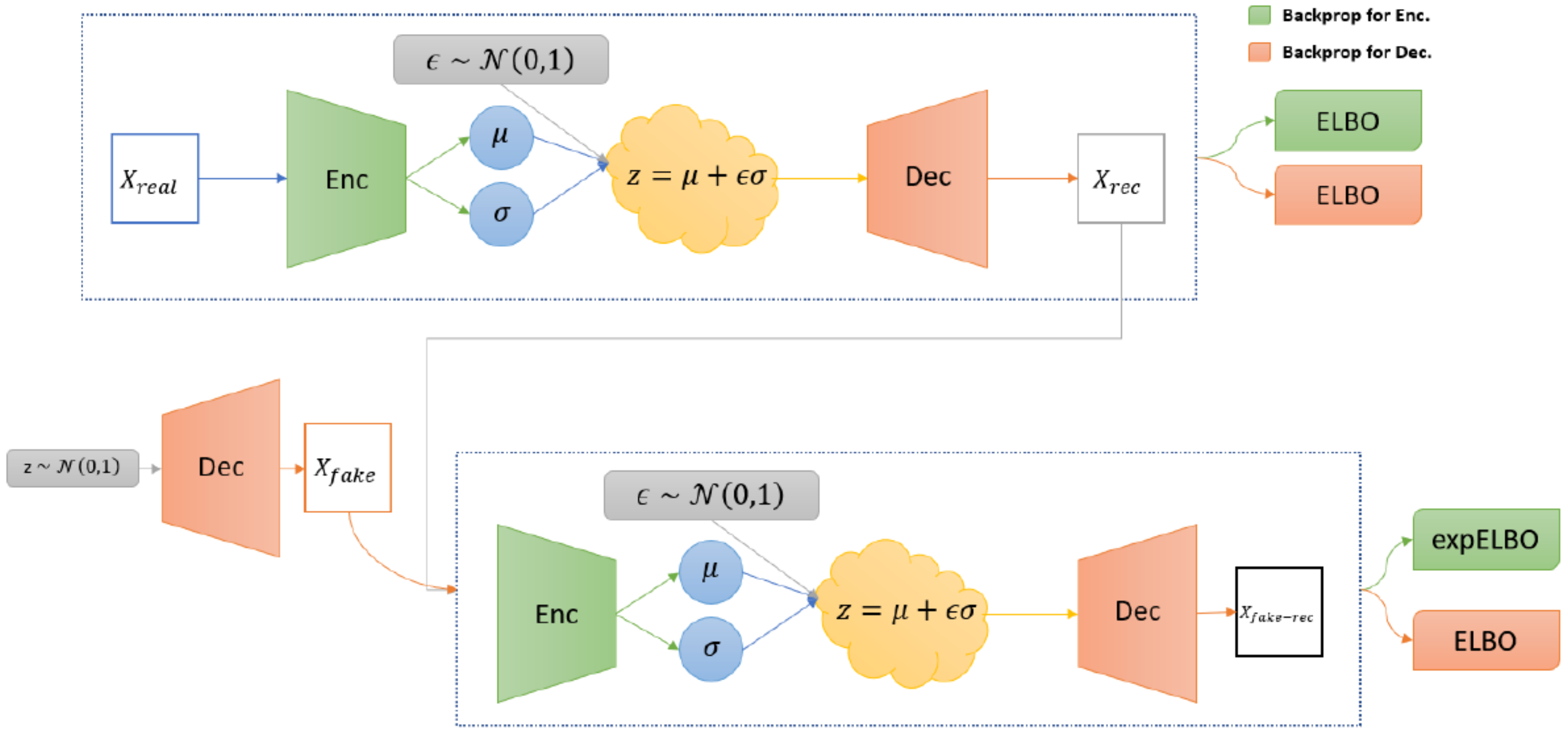
Soft-IntroVAE

- IntroVAE is hard to train
  - Can't reproduce
- Soft-IntroVAE
  - Utilizing the complete ELBO term instead of just the KL
  - Replacing the hard threshold with a soft exponential function over the ELBO

$$\mathcal{L}_{E_\phi}(x, z) = ELBO(x) - \frac{1}{\alpha}\exp(\alpha ELBO(D_\theta(z))),$$

$$\mathcal{L}_{D_\theta}(x, z) = ELBO(x) + \gamma ELBO(D_\theta(z)),$$

Soft-IntroVAE



$$\mathcal{L}_{E_\phi}(x, z) = ELBO(x) - \frac{1}{\alpha}\exp(\alpha ELBO(D_\theta(z))),$$

$$\mathcal{L}_{D_\theta}(x, z) = ELBO(x) + \gamma ELBO(D_\theta(z)),$$

Soft-IntroVAE

Algorithm 1 Training Soft-IntroVAE (pseudo-code)

Require: $\beta_{rec}, \beta_{kl}, \beta_{neg}, \gamma_r$

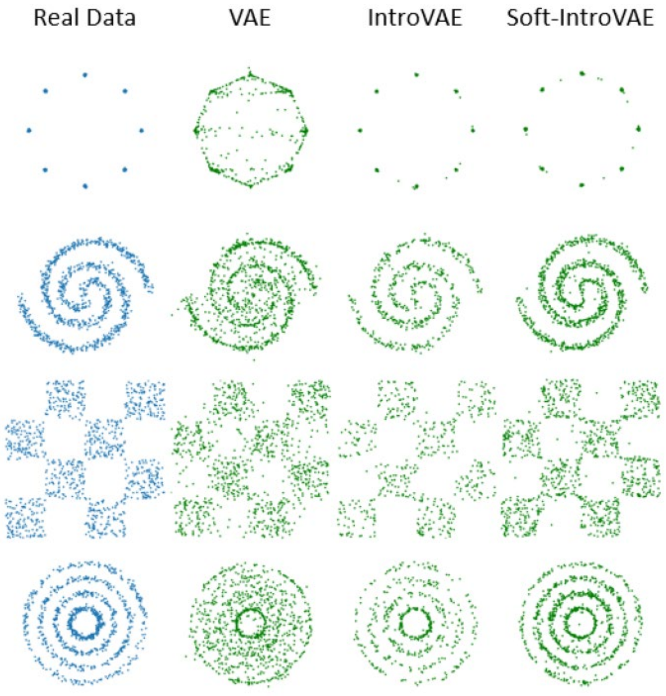
$\phi_E, \theta_D \leftarrow$ Initialize network parameters
$s \leftarrow 1/$input dim ▷ Scaling constant
while not converged **do**
  $X \leftarrow$ Random mini-batch from dataset
  $Z \leftarrow E(X)$ ▷ Encode
  $Z_f \leftarrow$ Samples from prior $N(0, I)$
  procedure UPDATEENCODER(ϕ_E)
    $X_r \leftarrow D(Z)$, $X_f \leftarrow D(Z_f)$ ▷ Decode
    $Z_{ff} \leftarrow E(X_f)$
    $X_{ff} \leftarrow D(Z_{ff})$
    ELBO $\leftarrow s \cdot ELBO(\beta_{rec}, \beta_{kl}, X, X_r, Z)$
    ELBO$_f \leftarrow ELBO(\beta_{rec}, \beta_{neg}, X_f, X_{ff}, Z_{ff})$
    expELBO$_f \leftarrow 0.5 \exp(2s \cdot \text{ELBO}_f)$
    $L_E \leftarrow$ ELBO $-$ expELBO$_f$ ▷ Eq. 4
    $\phi_E \leftarrow \phi_E + \eta \nabla_{\phi_E}(L_E)$ ▷ Adam update
  end procedure
  procedure UPDATEDECODER(θ_D)
    $X_r \leftarrow D(Z)$, $X_f \leftarrow D(Z_f)$ ▷ Decode
    $Z_{ff} \leftarrow E(X_f)$
    $X_{ff} \leftarrow sg(D(Z_{ff}))$ ▷ *sg: stop-gradient*
    ELBO $\leftarrow \beta_{rec} L_{rec}(X, X_r)$
    ELBO$_f \leftarrow ELBO(\gamma_r \cdot \beta_{rec}, \beta_{kl}, X_f, X_{ff}, Z_{ff})$
    $L_D \leftarrow s \cdot (\text{ELBO} + \text{ELBO}_f)$ ▷ Eq. 4
    $\theta_D \leftarrow \theta_D + \eta \nabla_{\theta_D}(L_D)$ ▷ Adam update
  end procedure
end while

Analysis

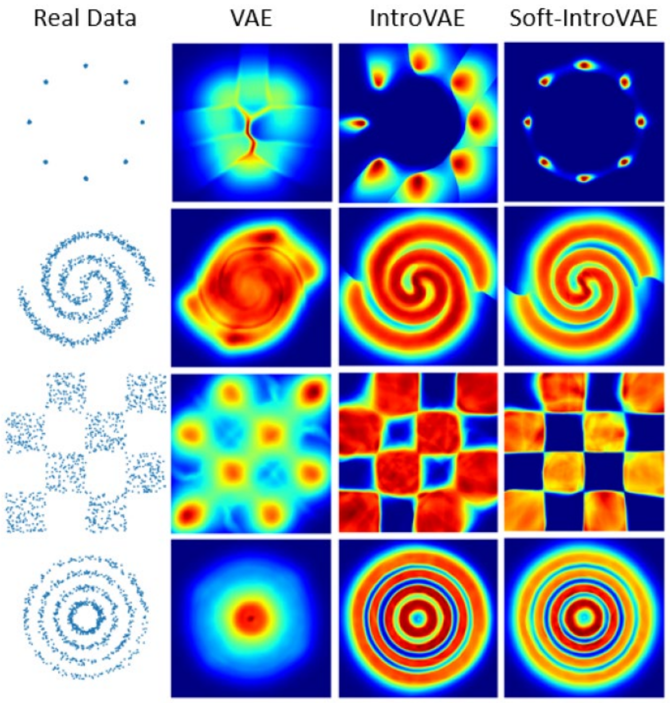
- Nash equilibrium
  - $$d^* \in \arg\min_d \{KL(p_{data} \| p_d) + \gamma H(p_d(x))\}.$$
  - No longer converging to p_data, but regularized by entropy
- Math is hard…

Experiments

- 2D Toy Dataset



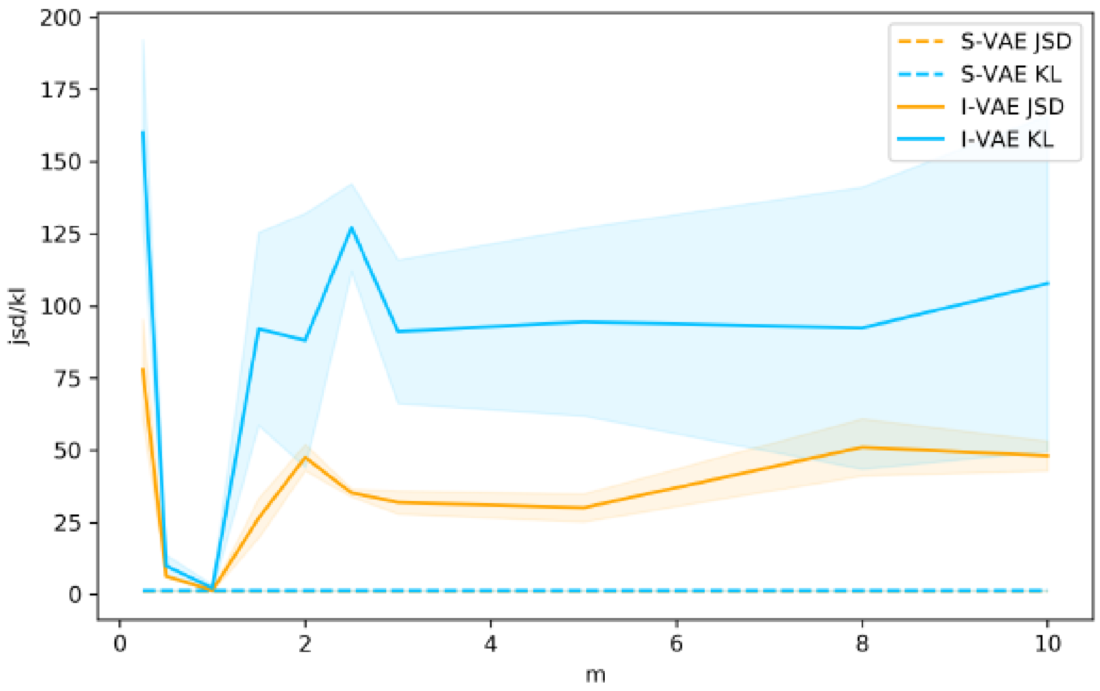
(a) Samples from the trained models.



(b) Density estimation with the trained models.

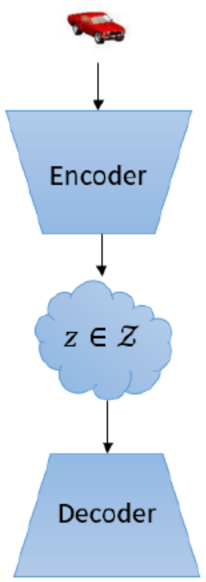
Experiments

- Training Stability
  - Probably due to the choice of *m* is sensitive

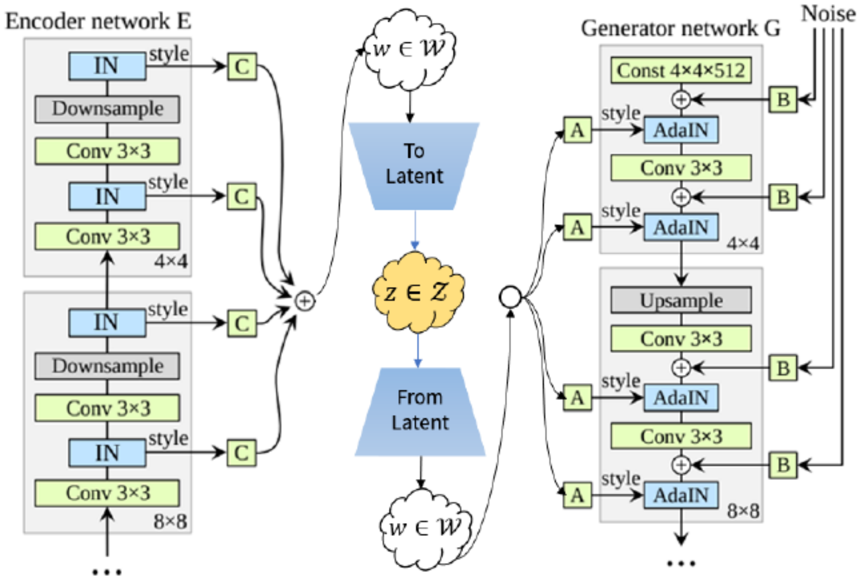


Experiments

- Image Generation
  - Architectures



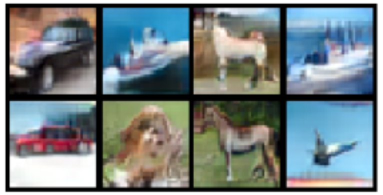
Standard Architecture



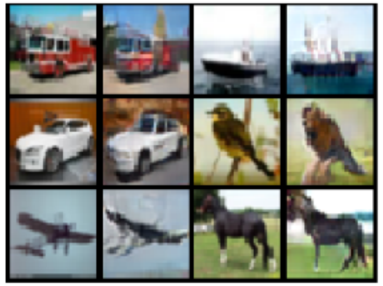
(c) Style-based Architecture inspired by [46]

Experiments

- Image Generation
  - Cifar-10



(a) Generated samples (FID: 4.6).

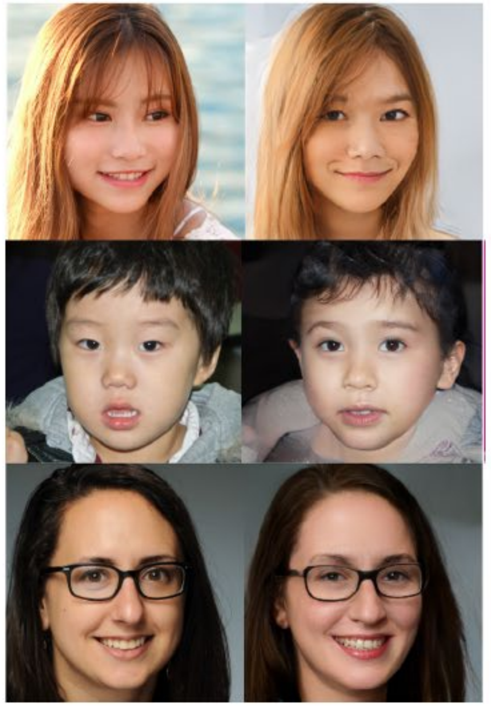


(b) Reconstructions on test data: Left: real, right: reconstruction.

Experiments

- Image Generation
  - CelebA-HQ and FFHQ datasets

(a) FFHQ dataset – samples from S-IntroVAE (FID: 17.55).



(b) FFHQ – reconstructions.

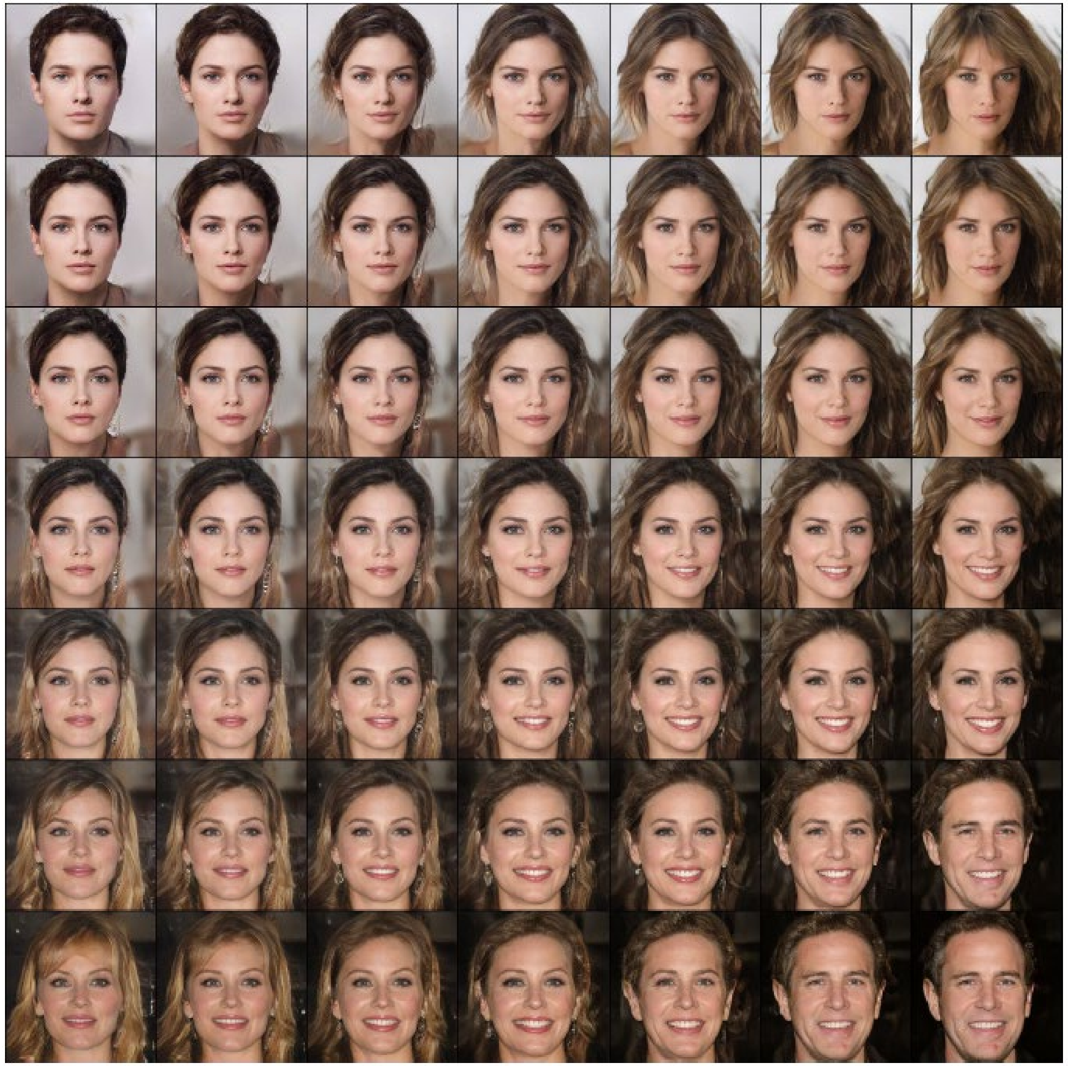
Experiments

- Image Generation
  - Interpolation in the latent space

Figure 5: Interpolation in the latent space between two samples from a model trained on CelebA-HQ.

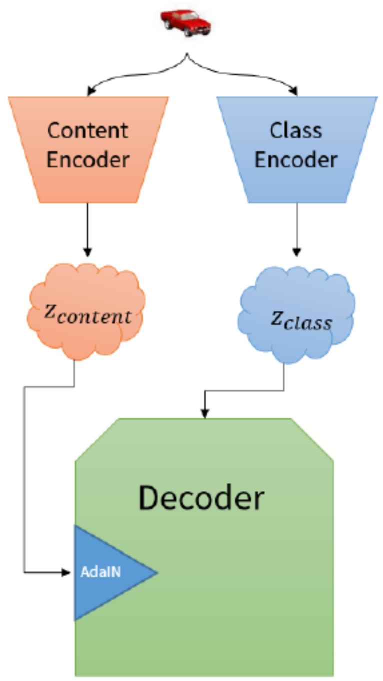
Experiments

- Image Generation
  - Interpolation in the latent space



Experiments

- Image Translation
  - Architectures



Disentanglement Architecture

Experiments

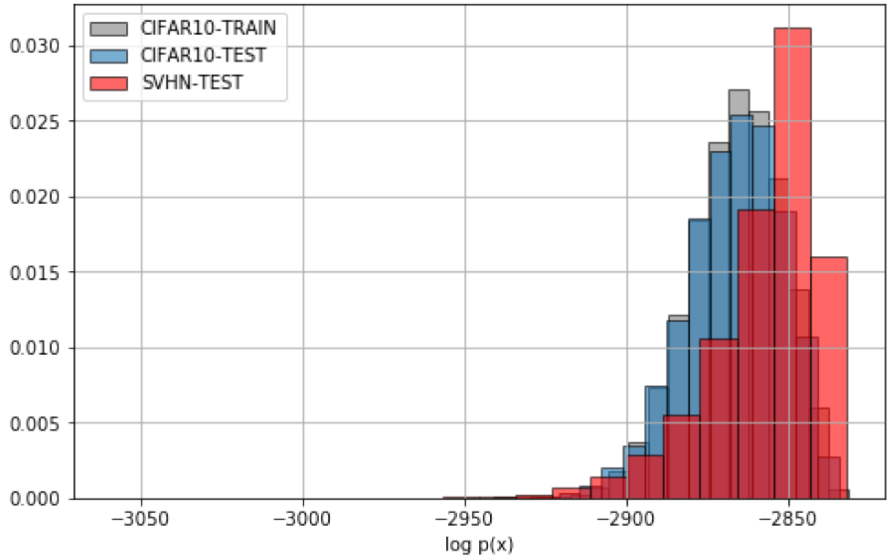
- Image Translation
  - Cars3D and KTH

Cars3D

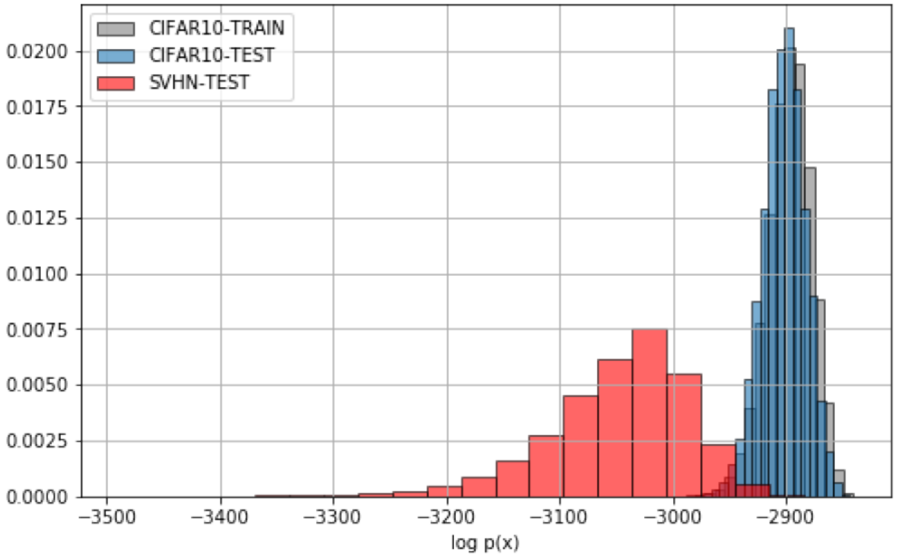
KTH

Experiments

- Out-of-Distribution (OOD) Detection
  - Cifar-10 & SVHN



(a) VAE



(b) Soft-IntroVAE

Conclusion

- Improve the training of IntroVAE
- A deeper theoretical understanding of IntroVAE

Thanks