paper ID: 2781

Disentangled Representation Learning for Controllable Image Synthesis: an Information-Theoretic Perspective

Shichang Tang
ShanghaiTech University
SIMIT, Chinese Academy of Sciences
University of Chinese Academy of Sciences

Xu Zhou
Shenzhen University

Xuming He
ShanghaiTech University

Yi Ma
University of California, Berkeley

Speaker: Shichang Tang

tangshch@shanghaitech.edu.cn

tsc2017@gmail.com

Main Contents

Introduction

Limitations of generative models

01 VAE and GAN have been developed for a few years，but we still do not know much about them

02 The latent space of GAN and VAE are usually uninterpretable, and thus uncontrolled

03 While GAN can synthesize better images, it does not have an encoder and cannot learn an latent representation given an input image

Goals

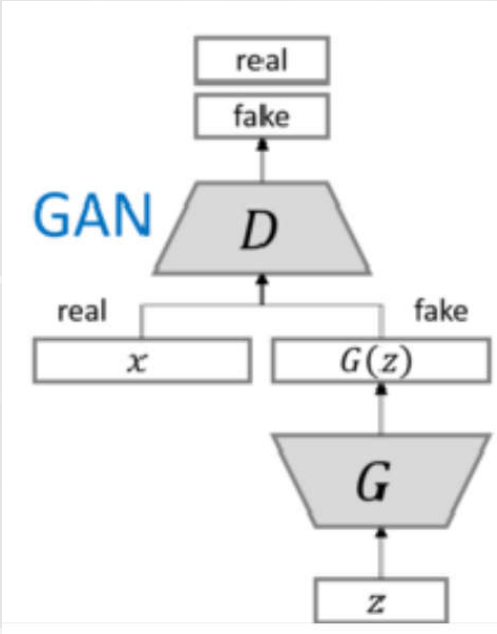
- ✓ Construct a flexible, interpretable generative model that can generate realistic images
- ✓ And be able to learn a disentangled representation so that the output images can be manipulated by controlling the latents

Background

Background

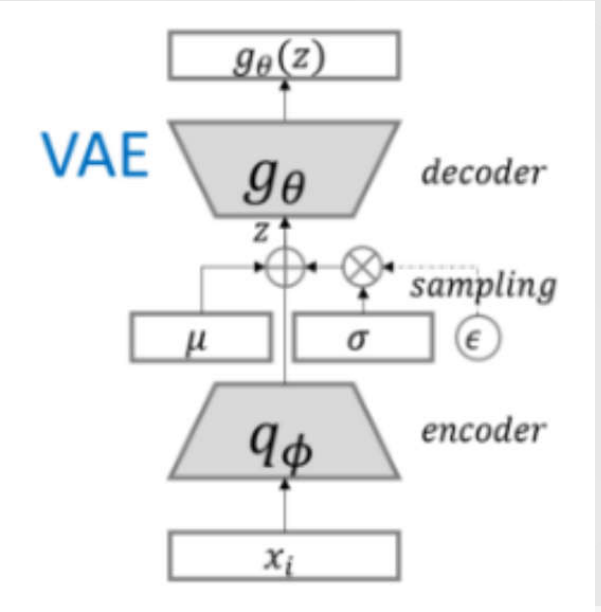
Generative Adversarial Nets (GAN)

GAN
WGAN
WGAN-GP
Info-GAN
LS-GAN
Style-GAN
BigGAN
...



Two Popular groups of Deep Generative Models

Variational AutoEncoder (VAE)



VAE
β-VAE
VQ-VAE
PixelVAE
...

Background: GAN

GAN

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_r}[\log D(x)] + \mathbb{E}_{z\sim P_z}[\log(1 - D(G(z)))]$$

WGAN

$$W(\mathbb{P}_r, P_g) = \min_G \max_D \mathbb{E}_{x\sim P_r}[D(x)] - \mathbb{E}_{\tilde{x}\sim P_g}[D(\tilde{x})]$$

where D is 1-Lipschitz

D tries to estimate a distance while G tries to minimize it

Different GANs

Info-GAN

$$I(c; G(z,c)) \geq \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c)$$

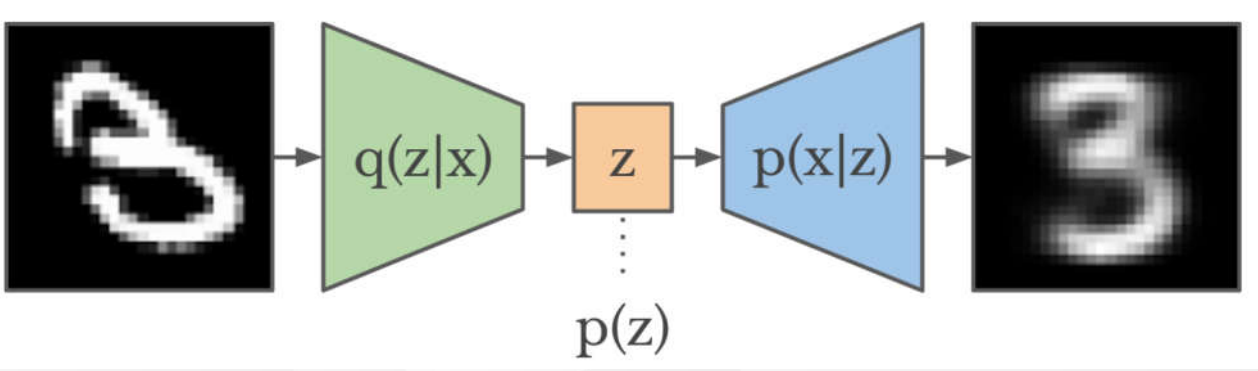
WGAN-GP

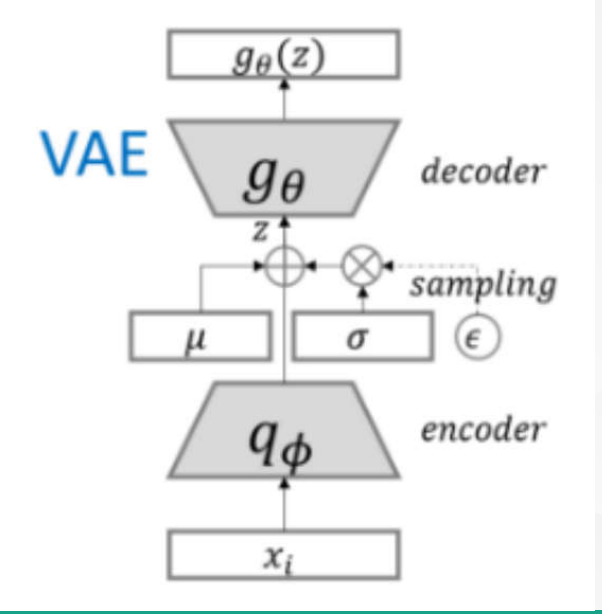
$$\mathbb{E}_{\tilde{x}\sim P_g}[D(\tilde{x})] - \mathbb{E}_{x\sim P_r}[D(x)] + \lambda\, \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}[(\left\|\nabla_{\hat{x}} D(\hat{x})\right\|_2 - 1)^2]$$

Background: VAE

Lower bound of log-likelihood:

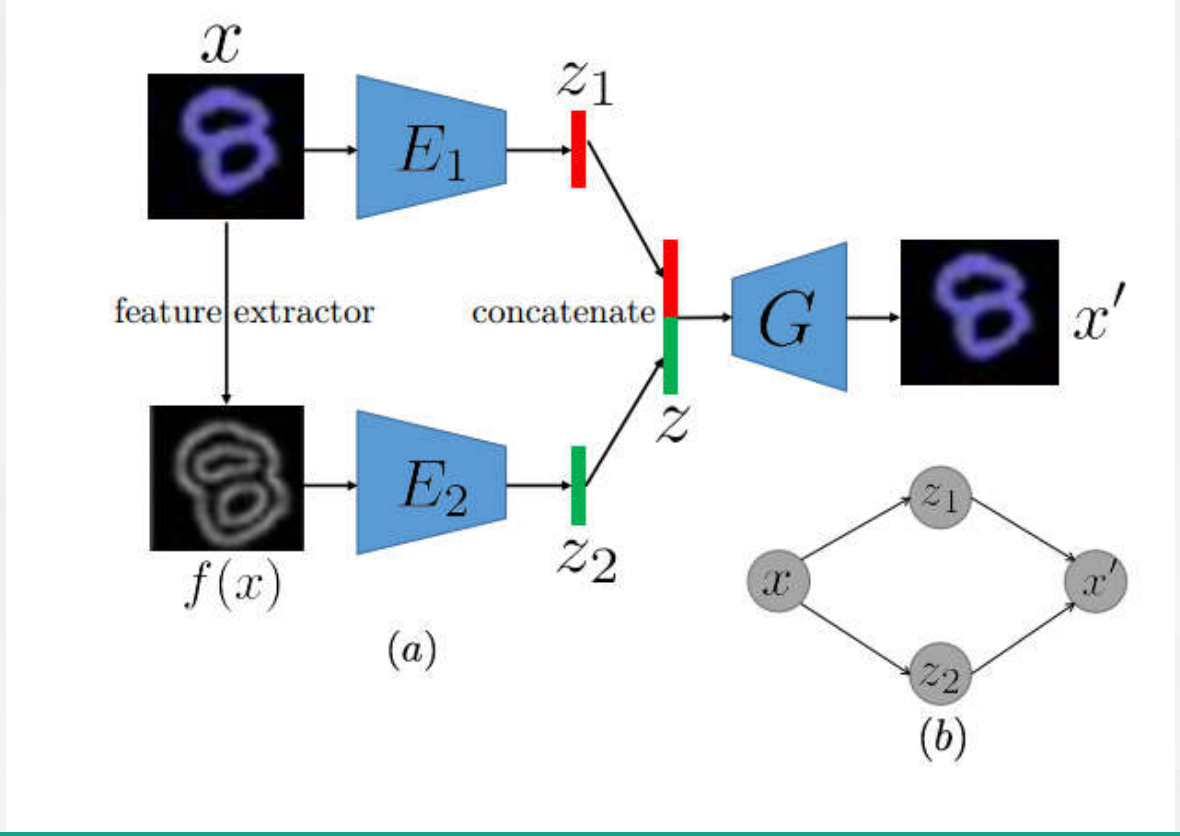
$$
\begin{aligned}
&\ln p(x) \\
&= \ln \int_z p(x,z)\,\mathrm{dz} \\
&= \ln \int_z p(x,z)\frac{q(z|x)}{q(z|x)}\,\mathrm{dz} \\
&\geq \mathbb{E}_{q(z|x)}[\ln \frac{p(x,z)}{q(z|x)}] \\
&= \mathbb{E}_{q(z|x)}[\ln \frac{p(x|z)p(z)}{q(z|x)}] \\
&= \mathbb{E}_{q(z|x)}[\ln p(x|z)] + \mathbb{E}_{q(z|x)}[\ln \frac{p(z)}{q(z|x)}] \\
&= \mathbb{E}_{q(z|x)}[\ln p(x|z)] + \int_z q(z|x)\ln \frac{p(z)}{q(z|x)}\,\mathrm{dz} \\
&= \mathbb{E}_{q(z|x)}[\ln p(x|z)] - D_{KL}[q(z|x)||p(z)] \\
&= likelihood - KL
\end{aligned}
$$





Disentangled Representation Learning for Controllable Image Synthesis

VAE with partitioned latent code



VAE with partitioned latent code

Different uses of generative models:

1.Feature learning:

Auto-encoding: $x \to (z_1, z_2) \to x'$

2.Generate images from scratch:

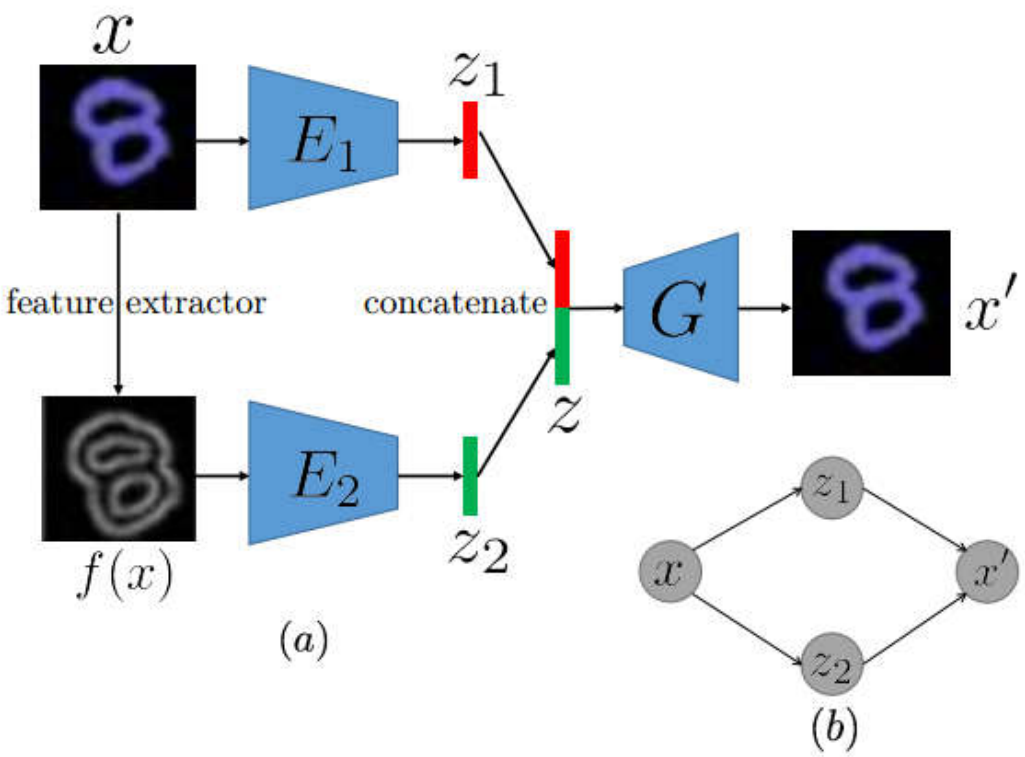
Sampling: $(\bar{z}_1, \bar{z}_2) \to \bar{x}$

3.Fusion (editing) :

Sample images independently: $x \sim P_r,\ \tilde{x} \sim P_r$

Encode: $z_1 \sim Q_1(\cdot|x)$ $\tilde{z}_2 \sim Q_2(\cdot|\tilde{x})$

Fusion: $(z_1, \tilde{z}_2) \to \hat{x}$



Entanglement of information

mutual information between z_2 and x' can be shared by z_1 :

$$I(z_2; x') = I(z_2; x'|z_1) + I(z_1; z_2; x').$$

e.g.,:

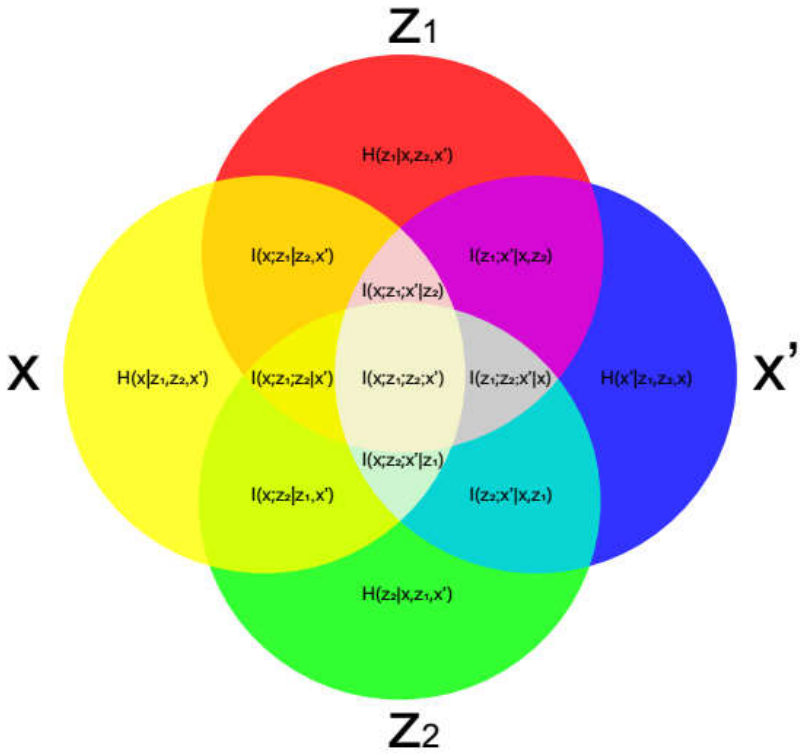
$$z_1 = E_1(x), \quad z_2 = E_2(x), \quad x' = E_1^{-1}(z_1)$$

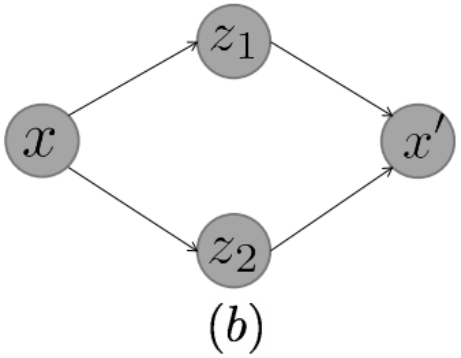
mutual information between z_2 and x' is maximized， but z_2 is not utilized by G at all

Therefore, mutual information between z_2 and x' should be maxminzed， but not shared by z_1 !

Approach： maximize the conditional mutual inforamtion：

$$I(z_2; x'|z_1)$$





VAE with partitioned latent code

Optimizable lower bounds: maximizing the conditional MI is to minimize the reconstruction loss when the latents are independent

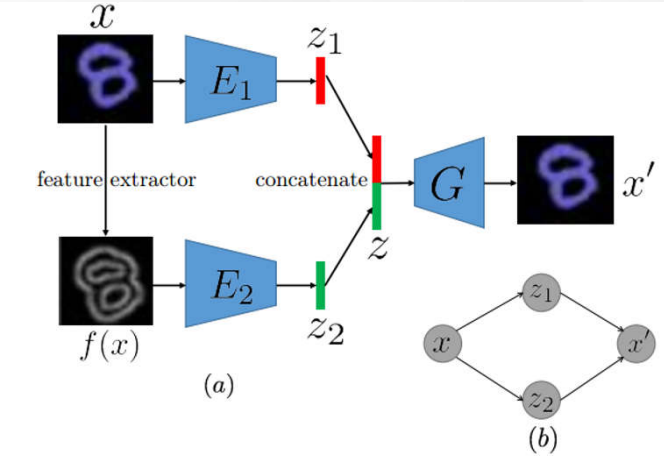
Maximize:

$$I(\tilde{z}_2; \hat{x}|z_1) \geq \mathbb{E}[\log q(\tilde{z}_2|z_1, \hat{x})] + H(\tilde{z}_2|z_1)$$

If $\tilde{z}_2$ and z_1 are independent，and $H(\tilde{z}_2)$ is constant, then one can minimize

$$L_{f1} = -\mathbb{E}[\log q(\tilde{z}_2|z_1, \hat{x})]$$ （Reconstruction loss of $\tilde{z}_2$）

If $Q(\cdot|z_1, \hat{x})$ is a Benoulli distribution, it becomes the cross-entropy loss



Or maximize

$$I(\tilde{z}_2; \hat{x}|z_1) \geq \mathbb{E}[\log q^*(f(\tilde{x})|z_1, \hat{x})] + H(f(x))$$ （Reconstruction loss of $f(\tilde{x})$）

Under a Gaussian assumption, we have

$$L_{f2} = \mathbb{E}[\|f(G(z_1, \tilde{z}_2)) - f(\tilde{x})\|_2^2]$$

The use of CGAN

Conditional mutual information maximization does not guarantee fidelity

e.g., samples overfitting the classifier can be unrealistic

D estimaites the Jensen-Shannon Divergence（JS-Divergence）between $P(\tilde{z}_2, \hat{x})$ 和 $P(z_2, x')$

$$L_D = -2JSD(P(z_2, x')\|P(\tilde{z}_2, \hat{x})) + \log(4)$$
$$= \mathbb{E}[\log D(z_2, x') + \log(1 - D(\tilde{z}_2, \hat{x}))]$$

While G tries to minimize：

$$L_{GAN} = -\mathbb{E}[\log(D(\tilde{z}_2, \hat{x}))]$$

Fusing color and sketch of CMNIST

Comparison methods:

VAE

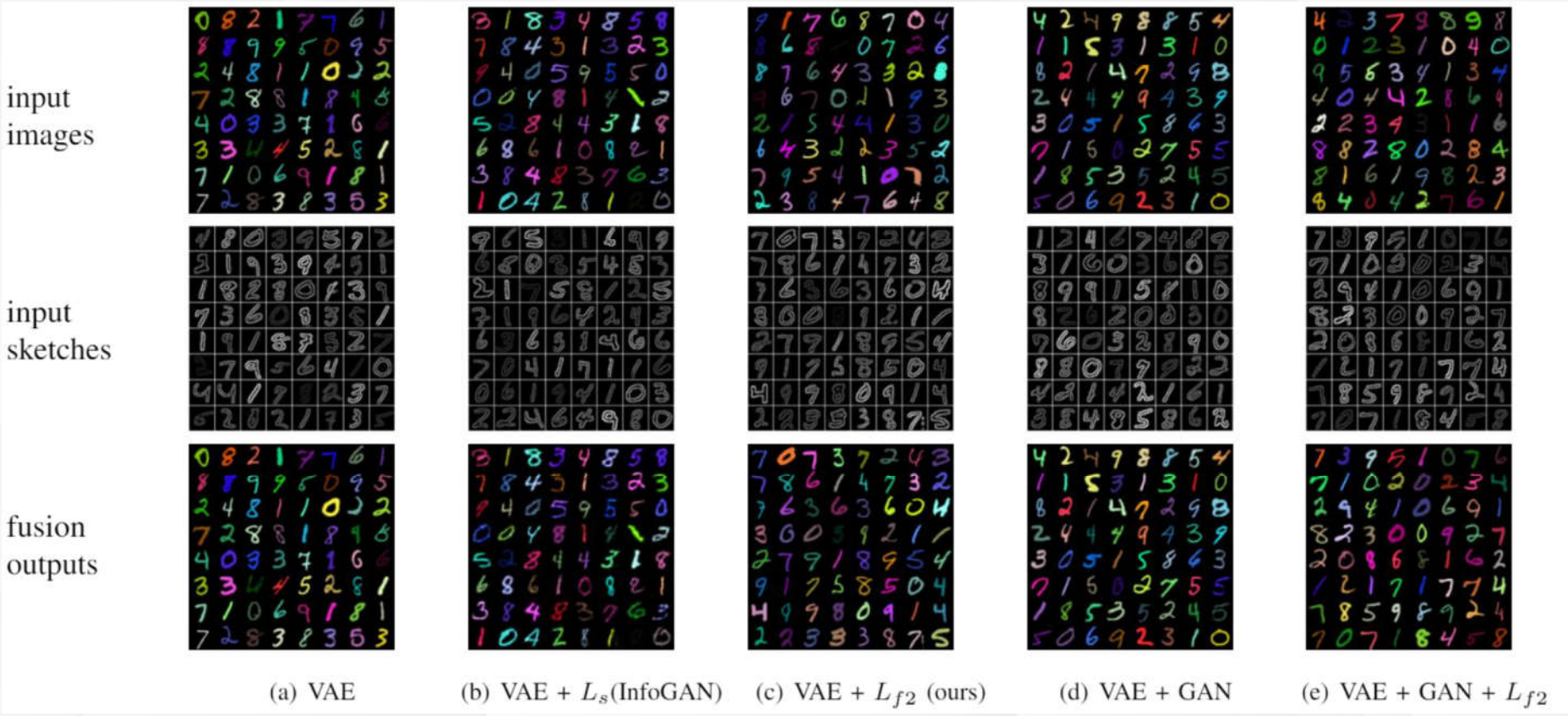
VAE+Ls(InfoGAN)

VAE+GAN

VAE+L_{f2} $$L_{f2} = \mathbb{E}[\|f(G(z_1, \tilde{z}_2)) - f(\tilde{x})\|_2^2]$$

VAE+GAN+L_{f2}

Fusing color and sketch of CMNIST



Controlling attributes of CelebA

Comparison methods:

CVAE $\log p(x|c) \geq \mathbb{E}_{z\sim Q(\cdot|x,c)}[\log p(x|z,c)] - D_{KL}(Q(\cdot|x,c)||P(z|c))$

VAE+GAN (Mathieu et al, also ablation study)

VAE+L_{f1}(ablation study) $L_{f1} = -\mathbb{E}[\log q(\tilde{z}_2|z_1,\hat{x})]$

VAE+GAN+L_{f1}(ours)

Flipping attributes of CelebA

Qualitative comparison:

Without the GAN loss, the output can be unrealistic; without the information loss, the output can be unchanged

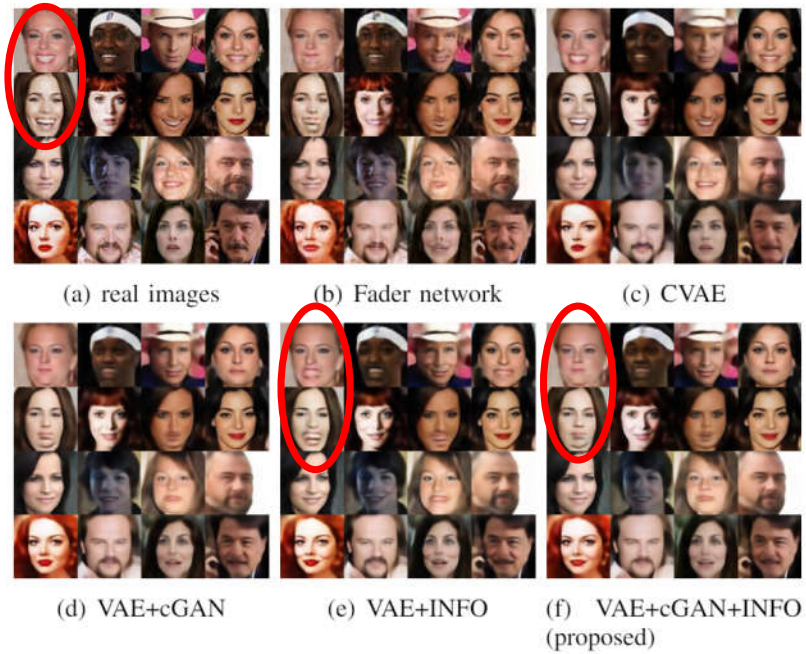


Fig. 4. Flipping z_2 when it represents "Smiling".

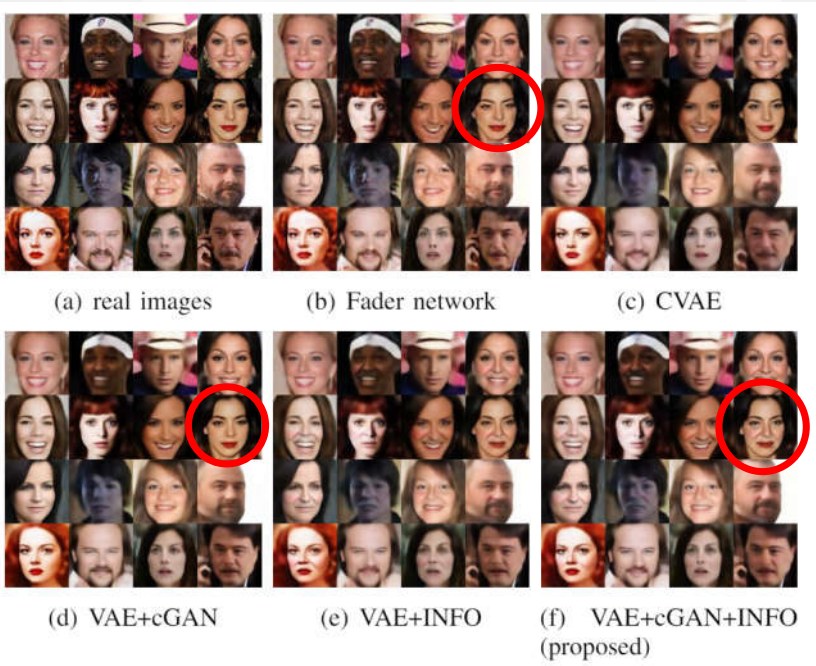


Fig. 5. Flipping z_2 when it represents "Young".

Controlling attributes of CelebA

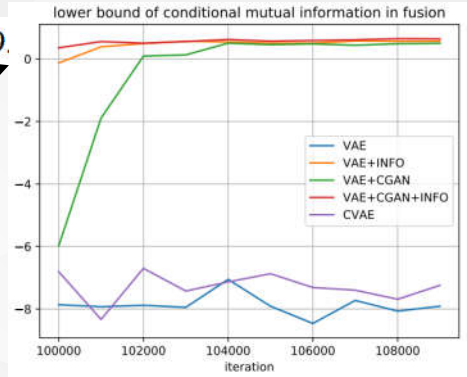
Quantitative comparisons:

$$I(\tilde{z}_2; \hat{x}|z_1) \leq H(\tilde{z}_2|z_1) = H(\tilde{z}_2) \approx 0.69$$

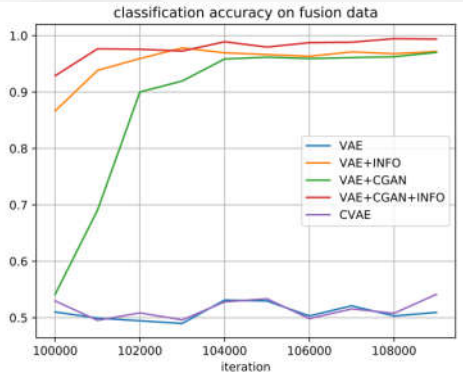
The lower bound approximates the upper bound

Classifier accuracy on fusion images

Lower bound of conditional mutual information

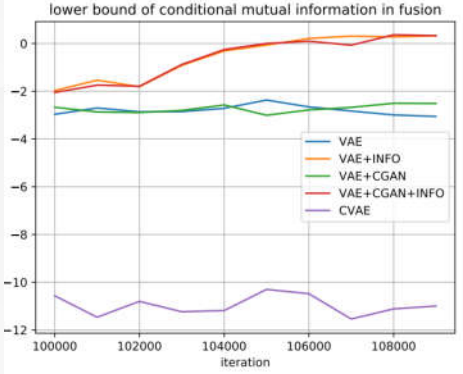


(a)

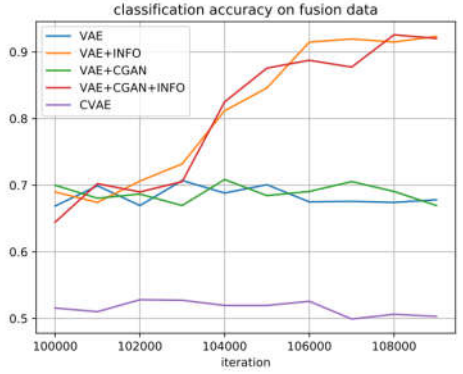


(b)

Smiling



(c)

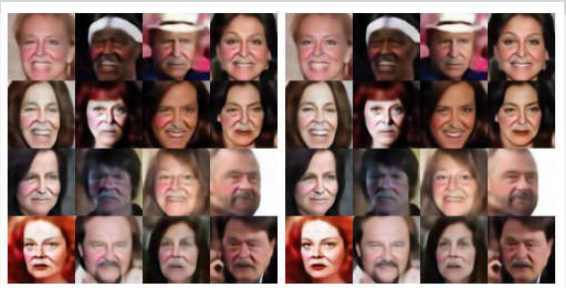


(d)

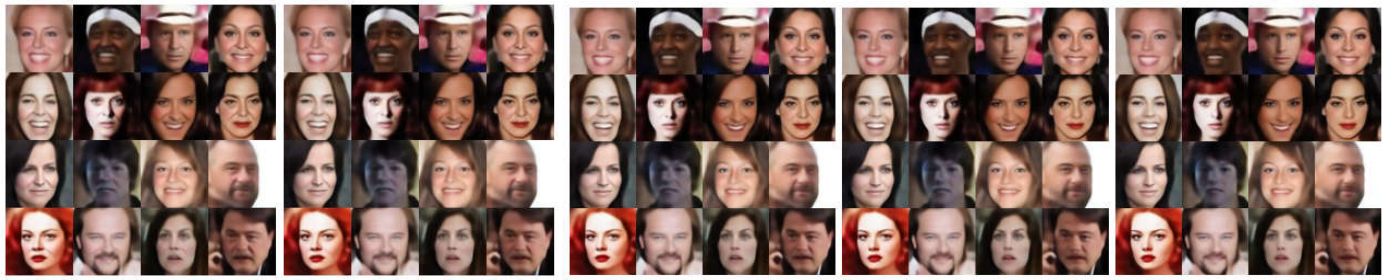
Young

Controlling attributes of CelebA

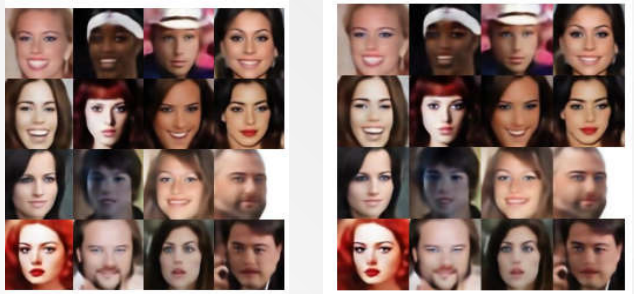
Interpolations and extrapolations (when z_2 represents "young"):



(a) $z_2 = -2$ (b) $z_2 = -1$



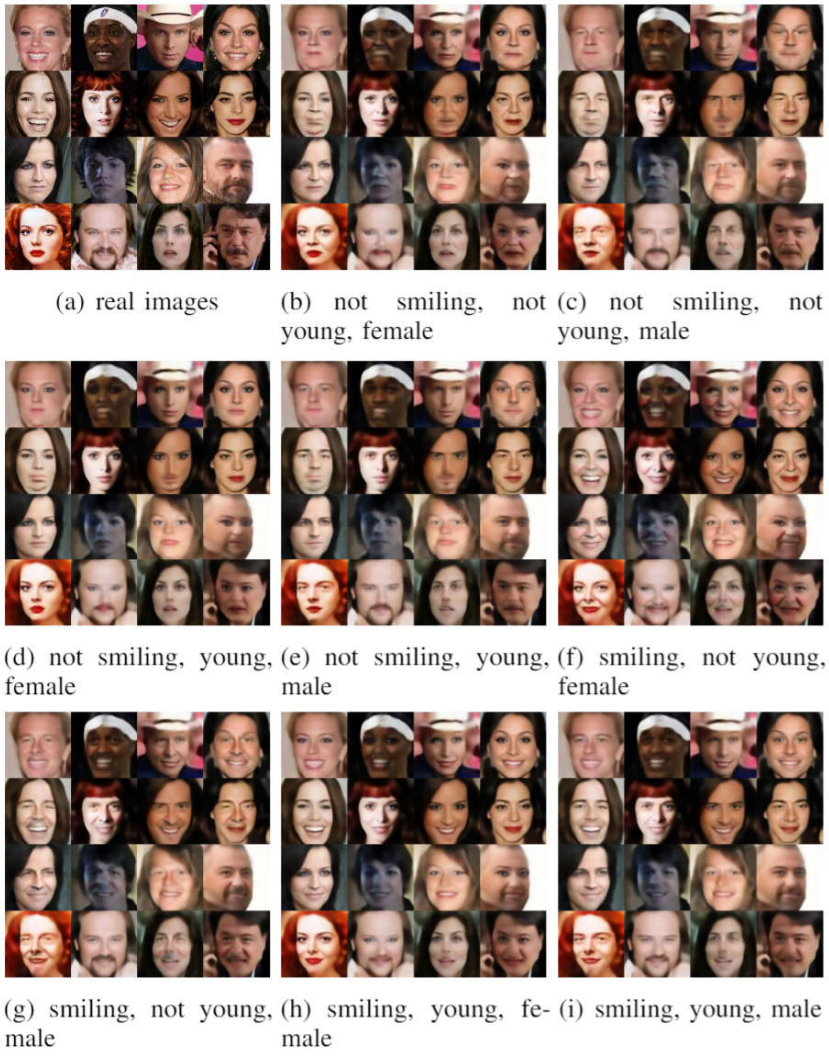
(a) $z_2 = 0$ (b) $z_2 = 0.25$ (c) $z_2 = 0.5$ (d) $z_2 = 0.75$ (e) $z_2 = 1$



(e) $z_2 = 2$ (f) $z_2 = 3$

Controlling attributes of CelebA

Control multiple attributes at the same time:



(a) real images (b) not smiling, not young, female (c) not smiling, not young, male

(d) not smiling, young, female (e) not smiling, young, male (f) smiling, not young, female

(g) smiling, not young, male (h) smiling, young, female (i) smiling, young, male

Conclusions

Conclusions

- We give a rigorous derivation of a variant of VAE with partitioned code, and analyze the phenomenon of entangled information from an information-theoretic perspective
- Derive optimizable lower bounds of conditional mutual information
- Use the proposed method to learn disentangled representation and perform controllable image synthesis

Future Work

- Use some recent state-of-the-art GANs such as StyleGAN2-ADA or MIX-GAN
- Resolve the difficulty in optimization when there are too many loss terms
- Experiment on some more attributions and ethnically diverse datasets

For more theoretical and experimental results, please refer to the paper and the supplementary material!

Thanks!