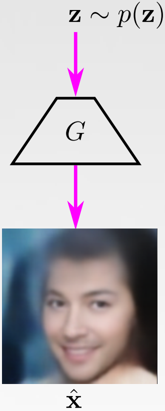
Laplacian Pyramid of Conditional Variational Autoencoders

Garoe Dorta[1,2] Sara Vicente[2] Lourdes Agapito[3]
Neill D.F. Campbell[1] Simon Prince[2] Ivor Simpson[2]

[1]University of Bath [2]Anthropics Technology Ltd. [3]University College London

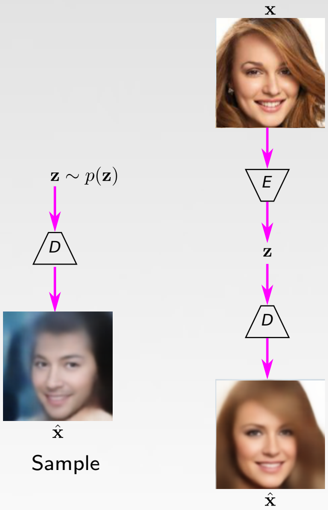
Generative models

Sampling and editing with generative models [1, 2, 3]



$\hat{\mathbf{x}}$	Sample
$\mathbf{z}$	Latent vector
$\boldsymbol{\theta}$	Model parameters
$G(\mathbf{z}; \boldsymbol{\theta})$	Generative function
$p(\mathbf{z})$	Simple known distribution

VAE

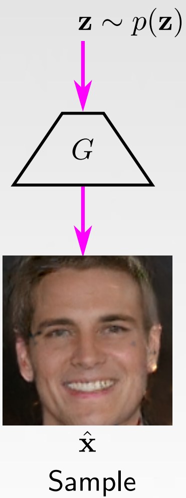


Sample

Reconstruction

Variational Autoencoders [4, 5]

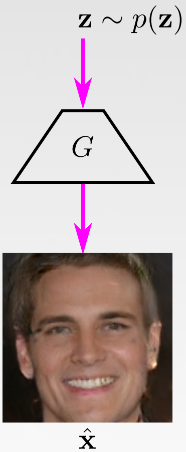
- Encoder $\mathbf{z} \sim \mathrm{E}(\mathbf{x})$
- Decoder $\hat{\mathbf{x}} \sim \mathrm{D}(\mathbf{z})$
- Gaussian likelihood estimation
  - Easy to train
  - Generate blurry images



Generative Adversarial Networks [6]

- Generator $\hat{\mathbf{x}} = G(\mathbf{z})$
- Discriminator
- Implicit likelihood estimation
  - Impressive results
  - Unstable training
- Only for sampling

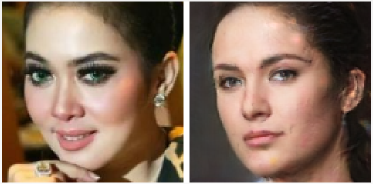
GAN



Sample

Generative Adversarial Networks [6]

- Generator $\hat{\mathbf{x}} = G(\mathbf{z})$
- Discriminator
- Implicit likelihood estimation
  - Impressive results
  - Unstable training
- Only for sampling

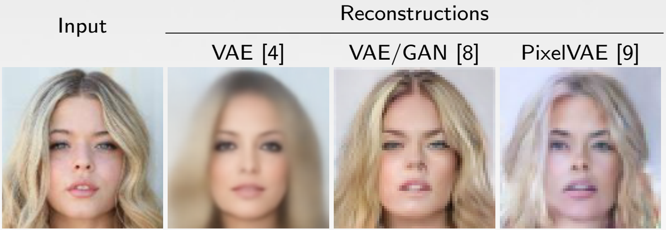


Input Reconstruction [7]

Related work

VAE extensions

- Complex distributions in the latent and output space [8, 9]
- Throw away the simplicity of the Gaussian likelihood estimation.

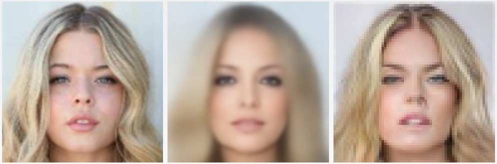


Related work

VAE extensions

- Complex distributions in the latent and output space [8, 9]
- Throw away the simplicity of the Gaussian likelihood estimation.

Input	Reconstructions		
	VAE [4]	VAE/GAN [8]	PixelVAE [9]

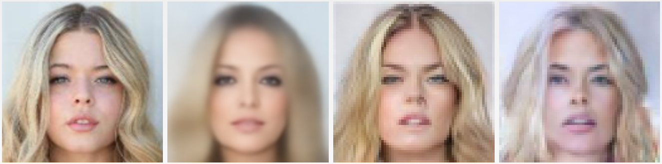


Related work

VAE extensions

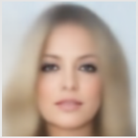
- Complex distributions in the latent and output space [8, 9]
- Throw away the simplicity of the Gaussian likelihood estimation.

Input	Reconstructions		
	VAE [4]	VAE/GAN [8]	PixelVAE [9]

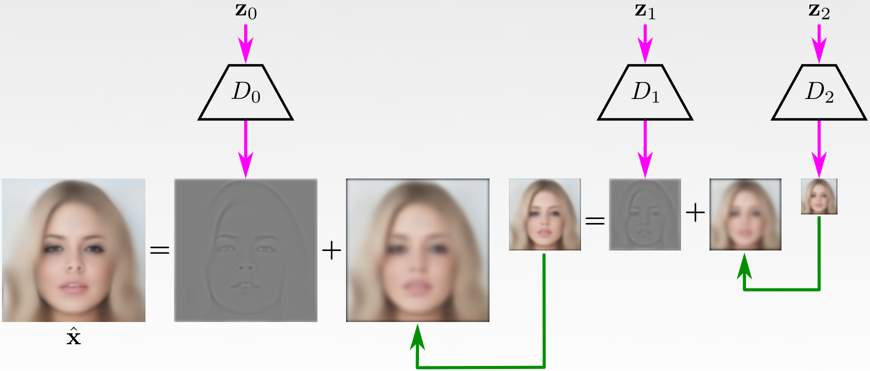


- Hierarchical approach
- Image generation in tractable steps
- Penalize errors in high-frequency images

VAE [4]



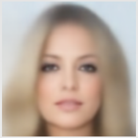
Ours



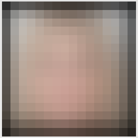
Our method

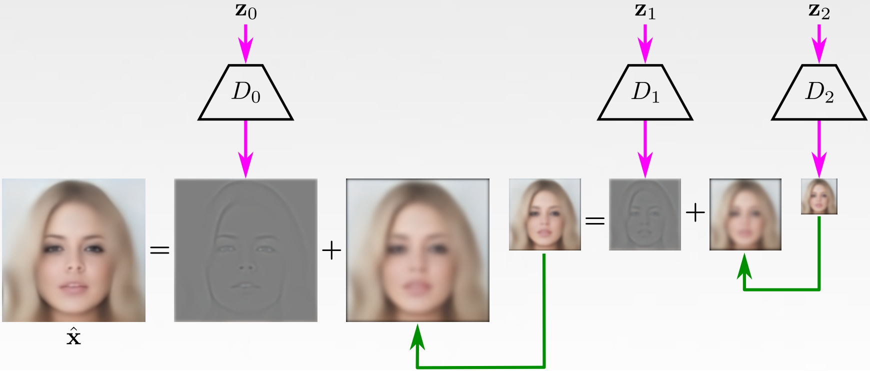
- Hierarchical approach
- Image generation in tractable steps
- Penalize errors in high-frequency images

VAE [4]

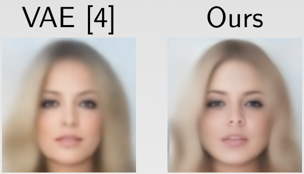


Ours





- Hierarchical approach
- Image generation in tractable steps
- Penalize errors in high-frequency images



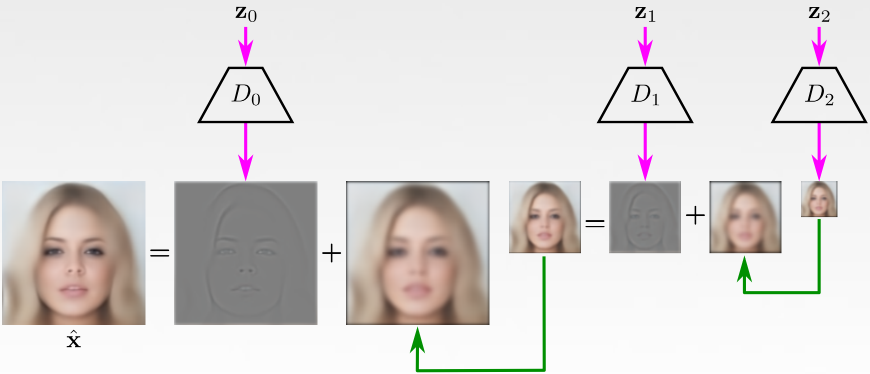


Image editing using the coarse-to-fine model structure

Reconstruction

Image editing using the coarse-to-fine model structure

Reconstruction

Coarse reconstruction

Image editing using the coarse-to-fine model structure

Reconstruction

Coarse painted

Our method: applications

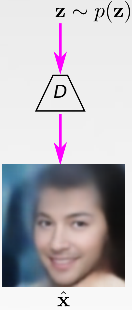
Image editing using the coarse-to-fine model structure

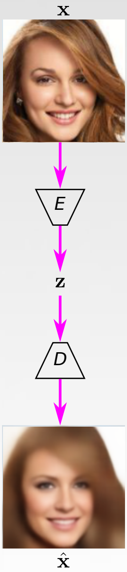
Reconstruction

Coarse painted

Edited

Background: VAE [4, 5]

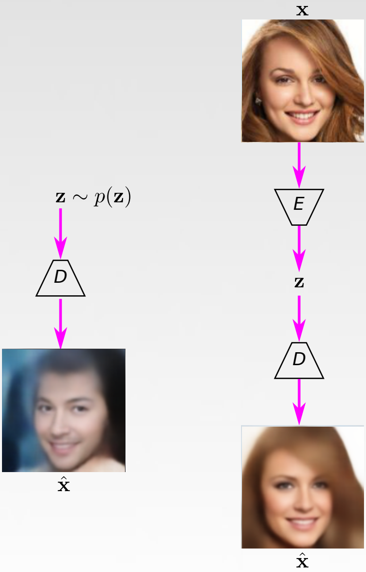




$$p(\mathbf{x} \,|\, \mathbf{z}; \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}\, \mathbf{I} \right)$$

$$\mathsf{D}(\mathbf{z}) = \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$$

Background: VAE [4, 5]

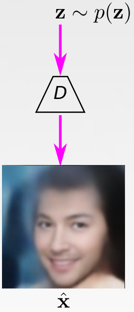


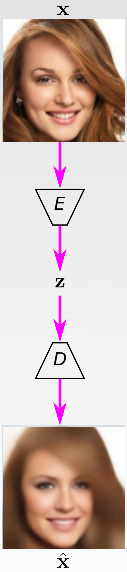
$$p(\mathbf{x} \,|\, \mathbf{z}; \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}\, \mathbf{I} \right)$$

$$q(\mathbf{z} \,|\, \mathbf{x}; \phi) = \mathcal{N}\left(\boldsymbol{\rho}, \boldsymbol{\omega}\, \mathbf{I} \right)$$

$$\mathsf{D}(\mathbf{z}) = \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}, \ \mathsf{E}(\mathbf{x}) = \{\boldsymbol{\rho}, \boldsymbol{\omega}\}$$

Background: VAE [4, 5]





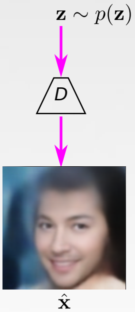
$$p(\mathbf{x} \,|\, \mathbf{z}; \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}\, \mathbf{I} \right)$$

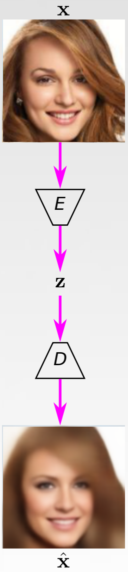
$$q(\mathbf{z} \,|\, \mathbf{x}; \phi) = \mathcal{N}\left(\boldsymbol{\rho}, \boldsymbol{\omega}\, \mathbf{I} \right)$$

$$\mathsf{D}(\mathbf{z}) = \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}, \;\; \mathsf{E}(\mathbf{x}) = \{\boldsymbol{\rho}, \boldsymbol{\omega}\}$$

$$L = \overbrace{-\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x};\phi)} \left[\log \; p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})\right]}^{\text{Reconstruction loss}} + \overbrace{D_{KL}\left[q(\mathbf{z}|\mathbf{x};\phi) || p(\mathbf{z})\right]}^{\text{Latent space loss}}$$

Background: VAE [4, 5]





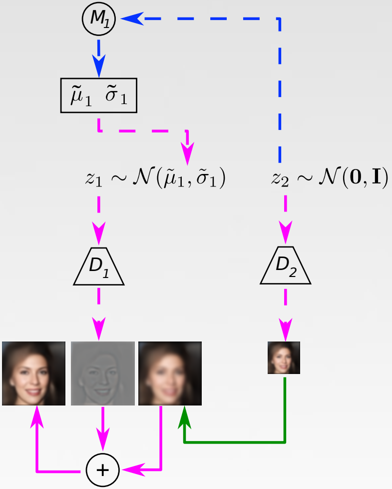
$$p(\mathbf{x} \,|\, \mathbf{z}; \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\sigma}\, \mathbf{I} \right)$$

$$q(\mathbf{z} \,|\, \mathbf{x}; \boldsymbol{\phi}) = \mathcal{N}\left(\boldsymbol{\rho}, \boldsymbol{\omega}\, \mathbf{I} \right)$$

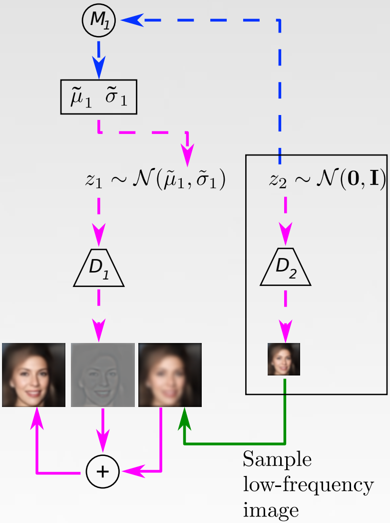
$$\mathsf{D}(\mathbf{z}) = \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}, \;\; \mathsf{E}(\mathbf{x}) = \{\boldsymbol{\rho}, \boldsymbol{\omega}\}$$

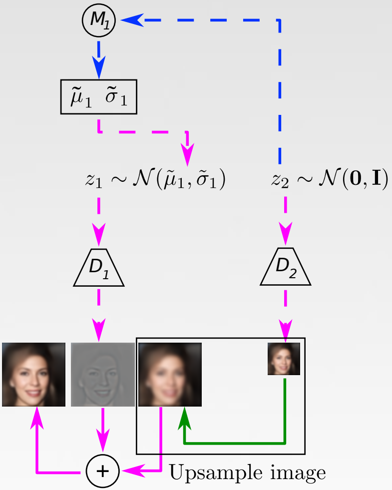
$$L = \overbrace{-\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x};\boldsymbol{\phi})} \left[\log \; p(\mathbf{x}|\mathbf{z}; \boldsymbol{\theta}) \right]}^{\text{Reconstruction loss}} + \overbrace{D_{KL} \left[q(\mathbf{z}|\mathbf{x}; \boldsymbol{\phi}) || p(\mathbf{z}) \right]}^{\text{Latent space loss}}$$

$$p(\mathbf{z}) = \overbrace{\mathcal{N}\left(\mathbf{0}, \mathbf{I} \right))}^{\text{Prior}}$$

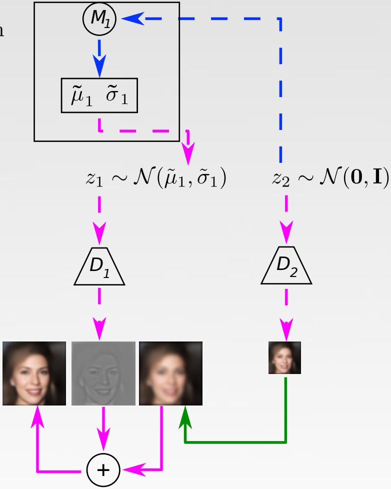


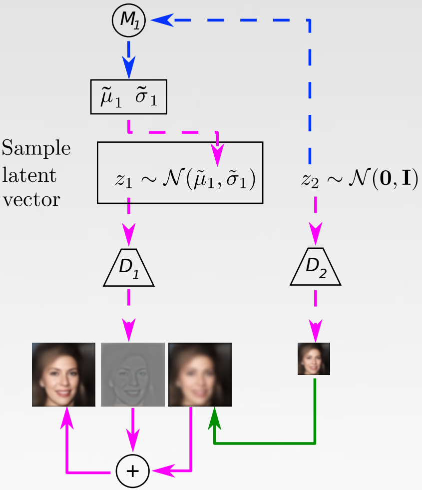
M_1
$\tilde{\mu}_1 \quad \tilde{\sigma}_1$
$z_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1)$
$z_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
D_1
D_2
+

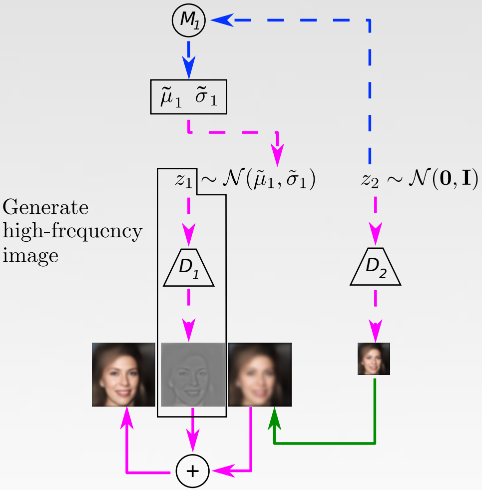

M_1
$\tilde{\mu}_1 \ \tilde{\sigma}_1$
$z_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1)$
$z_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
D_1
D_2
+
Sample
low-frequency
image

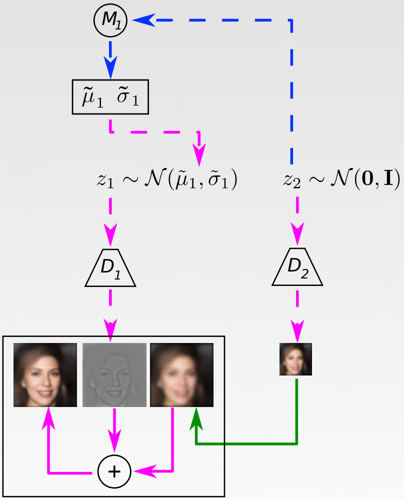

M_1
$\tilde{\mu}_1 \ \tilde{\sigma}_1$
$z_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1)$
$z_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
D_1
D_2
+
Upsample image

Predict the mean and variance of the latent vector

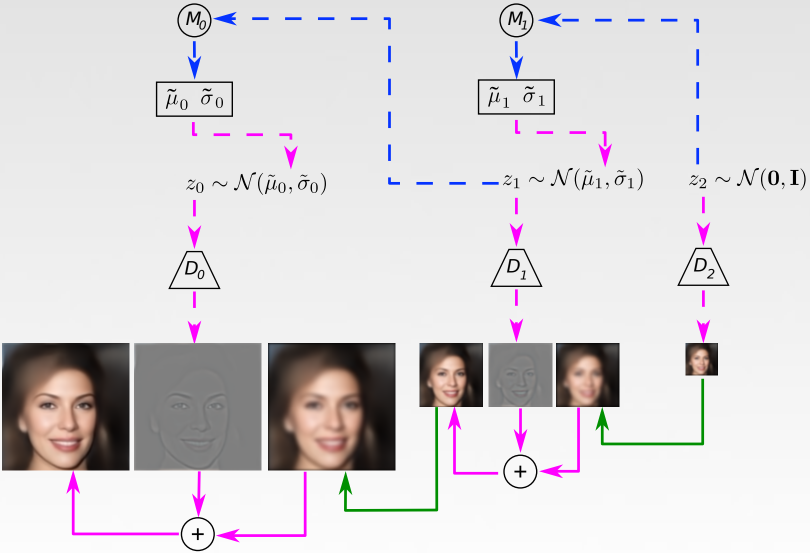



M_1
$\tilde{\mu}_1 \;\; \tilde{\sigma}_1$
Sample latent vector
$z_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1)$
$z_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
D_1
D_2
+

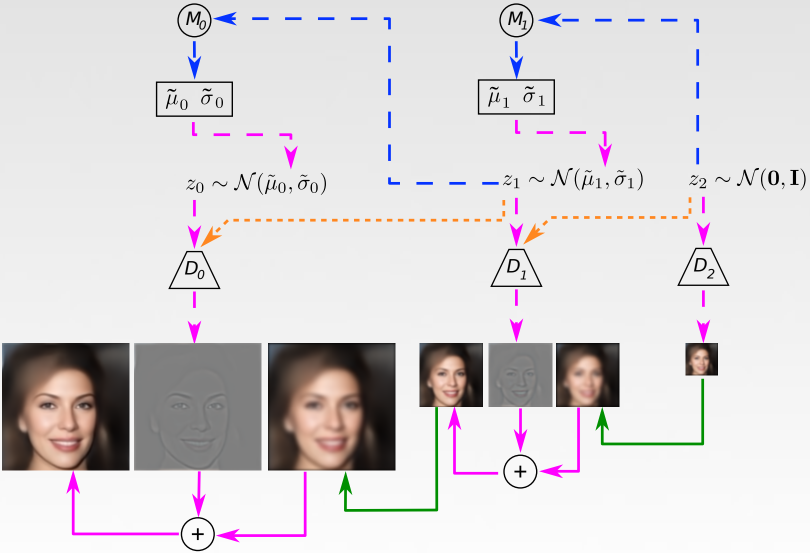

M_1
$\tilde{\mu}_1 \;\; \tilde{\sigma}_1$
$z_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1)$
$z_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Generate
high-frequency
image
D_1
D_2
+



Add high-frequency image to low-frequency image

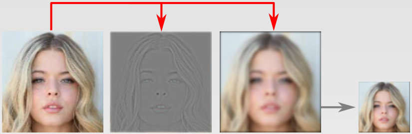


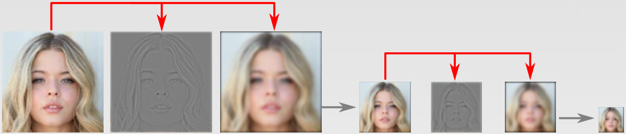
M0
M1
μ̃0 σ̃0
μ̃1 σ̃1
z0 ∼ N(μ̃0, σ̃0)
z1 ∼ N(μ̃1, σ̃1)
z2 ∼ N(0, I)
D0
D1
D2
+
+



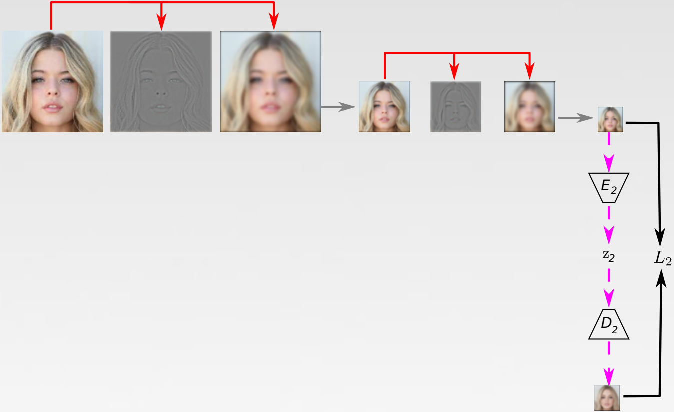
M_0
M_1
$\tilde{\mu}_0 \ \tilde{\sigma}_0$
$\tilde{\mu}_1 \ \tilde{\sigma}_1$
$z_0 \sim \mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0)$
$z_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1)$
$z_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
D_0
D_1
D_2
+
+

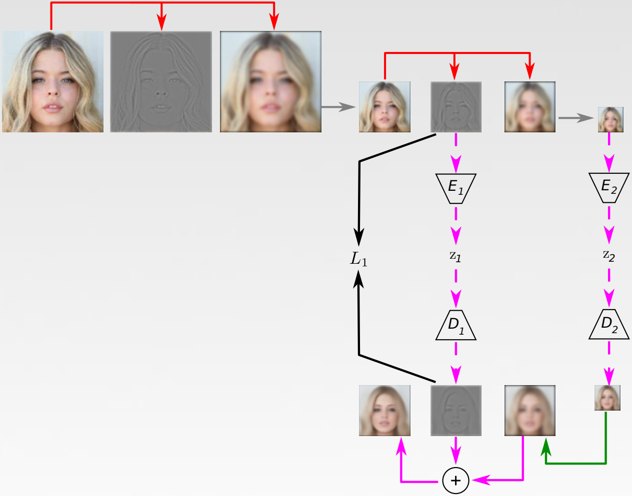
Methodology: Learning

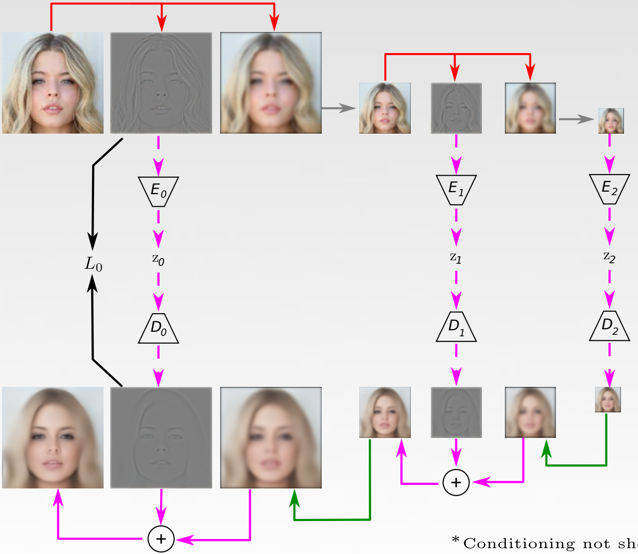




Methodology: Learning




E_1
E_2
L_1
z_1
z_2
D_1
D_2
+



*Conditioning not shown

Input image

$\mathbf{x}_k$

$=$

High frequency

$\mathbf{h}_k$

$+$

Low frequency

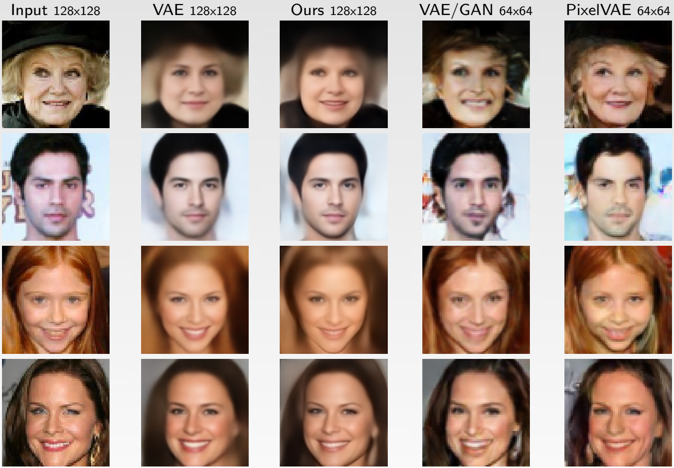
$\mathsf{u}(\mathbf{x}_{k+1})$

$$L_k = \overbrace{-\mathbb{E}_{\mathbf{z}_k \sim q_k(\mathbf{z}_k|\mathbf{h}_k, u(\mathbf{x}_k); \phi_k)} \left[\log \; p_k(\mathbf{h}_k|\mathbf{z}_k, \mathbf{z}_{k+1}, \cdots, \mathbf{z}_K; \boldsymbol{\theta}_k)\right]}^{\text{Reconstruction loss}} +$$

$$\overbrace{\lambda_k D_{KL}\left[q_k(\mathbf{z}_k|\mathbf{h}_k, u(\mathbf{x}_{k+1}); \phi_k)||p(\mathbf{z}_k)\right]}^{\text{Latent space loss}}$$

$$p(\mathbf{z}_k) = \overbrace{\mathcal{N}\left(R_k(\boldsymbol{\mu}_{k+1}; \boldsymbol{\xi}_k), S_k(\boldsymbol{\sigma}_{k+1}; \boldsymbol{\omega}_k)\right)}^{\text{Prior}}, \quad M_k = \overbrace{\{R_k, S_k\}}^{\text{Prior network}}$$

Reconstructions



Comparison of image reconstructions

Model	Error ($\sqrt{\text{MSE}}$)
VAE [4] 64×64	22.78 ± 4.64
VAE/GAN [8] 64×64	30.49 ± 7.32
Ours 64×64	$\mathbf{20.60 \pm 4.81}$
VAE [4] 128×128	20.75 ± 4.40
Ours 128×128	$\mathbf{20.61 \pm 5.15}$

Quantitative model comparison of image reconstructions

Reconstructions

Model	Error ($\sqrt{\text{MSE}}$)
VAE [4] 64×64	22.78 ± 4.64
VAE/GAN [8] 64×64	30.49 ± 7.32
Ours 64×64	$\mathbf{20.60} \pm \mathbf{4.81}$
VAE [4] 128×128	20.75 ± 4.40
Ours 128×128	$\mathbf{20.61} \pm \mathbf{5.15}$

Quantitative model comparison of image reconstructions

A

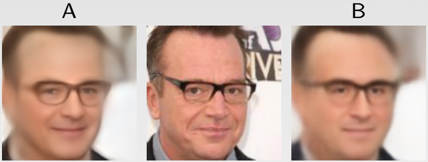
B

Model	Preference %
	Without original
VAE [4]	15.61 ± 8.14
Ours	$\mathbf{84.39} \pm \mathbf{8.14}$

User study: evaluation of pairs of reconstructions

Model	Error ($\sqrt{\text{MSE}}$)
VAE [4] 64×64	22.78 ± 4.64
VAE/GAN [8] 64×64	30.49 ± 7.32
Ours 64×64	$\mathbf{20.60 \pm 4.81}$
VAE [4] 128×128	20.75 ± 4.40
Ours 128×128	$\mathbf{20.61 \pm 5.15}$

Quantitative model comparison of image reconstructions



Model	Preference %	
	Without original	With original
VAE [4]	15.61 ± 8.14	26.30 ± 7.35
Ours	$\mathbf{84.39 \pm 8.14}$	$\mathbf{73.70 \pm 7.35}$

User study: evaluation of pairs of reconstructions

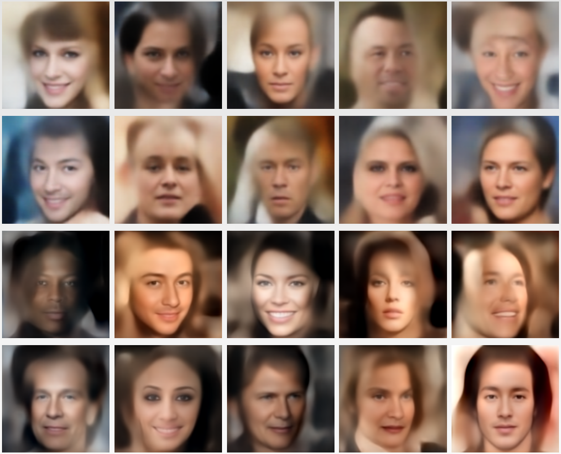
VAE

Samples from VAE and our model

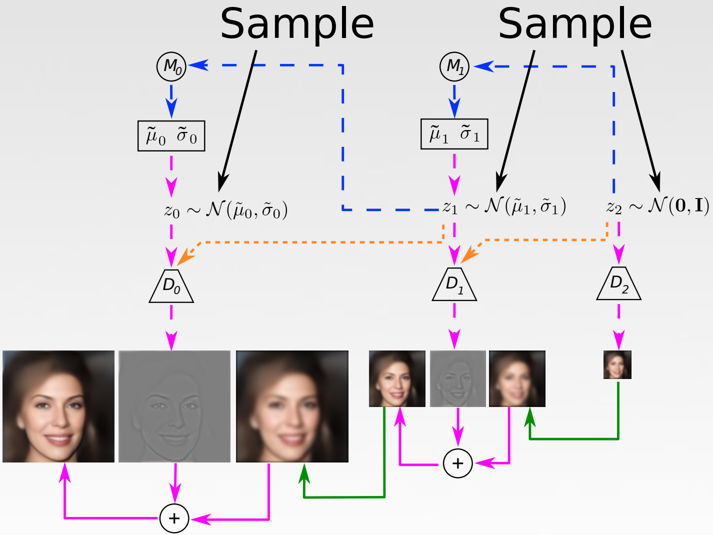
Samples

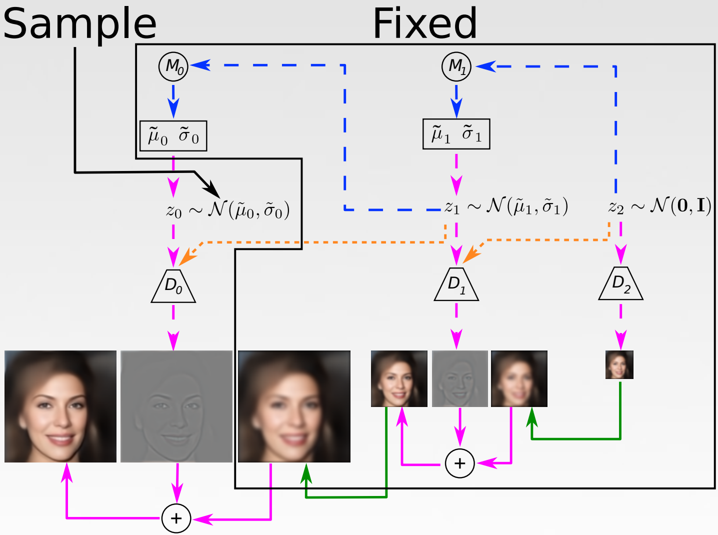
VAE

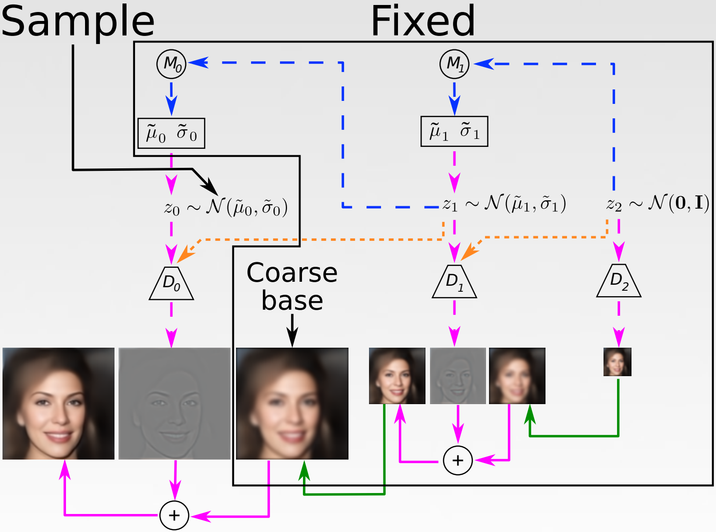
Ours

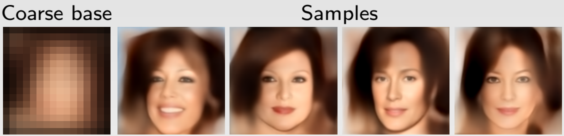


Samples from VAE and our model


Sample
Sample
M_0
M_1
$\tilde{\mu}_0 \;\; \tilde{\sigma}_0$
$\tilde{\mu}_1 \;\; \tilde{\sigma}_1$
$z_0 \sim \mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0)$
$z_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1)$
$z_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
D_0
D_1
D_2
+
+

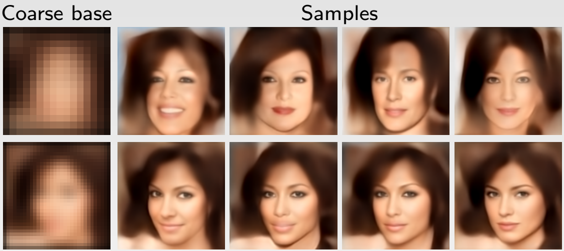

Sample
Fixed
M_0
M_1
$\tilde{\mu}_0 \;\; \tilde{\sigma}_0$
$\tilde{\mu}_1 \;\; \tilde{\sigma}_1$
$z_0 \sim \mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0)$
$z_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1)$
$z_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
D_0
D_1
D_2
+
+


Sample
Fixed
M_0
$\tilde{\mu}_0 \;\; \tilde{\sigma}_0$
$z_0 \sim \mathcal{N}(\tilde{\mu}_0, \tilde{\sigma}_0)$
D_0
M_1
$\tilde{\mu}_1 \;\; \tilde{\sigma}_1$
$z_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1)$
$z_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
D_1
D_2
Coarse base
+
+



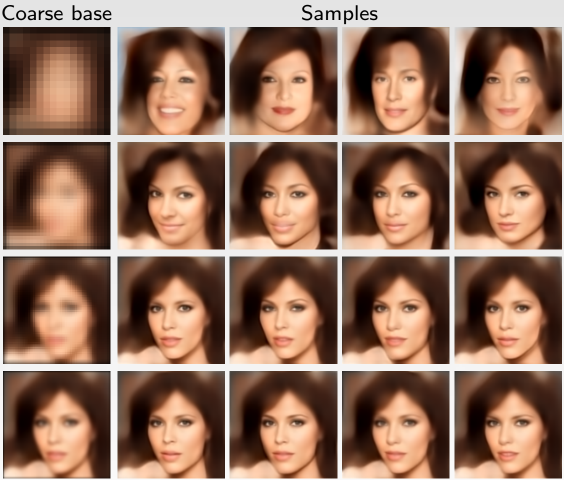
Sampling with $\mathbf{z}_{k \ldots K}$ fixed at different levels of the pyramid

Samples



Sampling with $\mathbf{z}_{k\ldots K}$ fixed at different levels of the pyramid

Samples



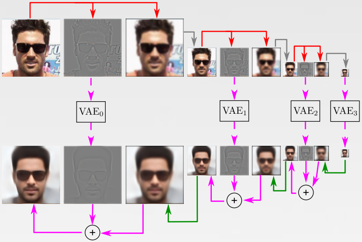
Sampling with $\mathbf{z}_{k\ldots K}$ fixed at different levels of the pyramid

Input

Laplacian pyramid of input

Input

Reconstructed



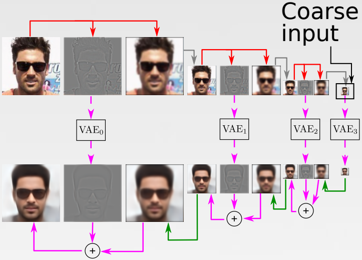
Reconstruct input

Editing: removing glasses

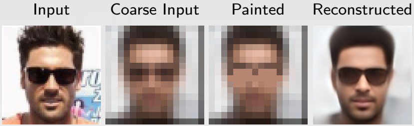
Input

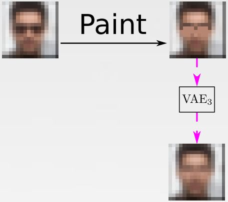
Coarse Input

Reconstructed



Select coarse level

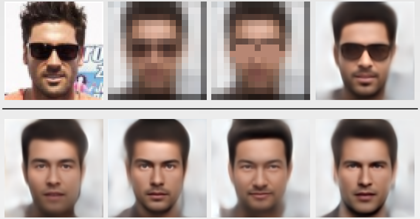


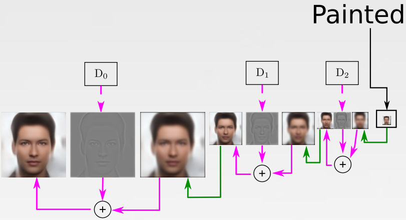


Paint coarse level

Editing: removing glasses

Input | Coarse Input | Painted | Reconstructed

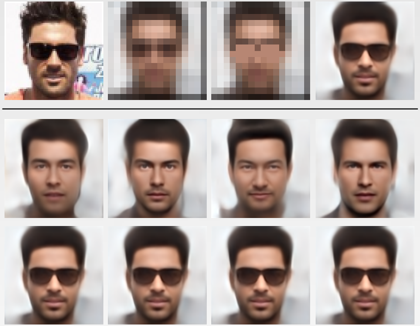




Sample from painted coarse image

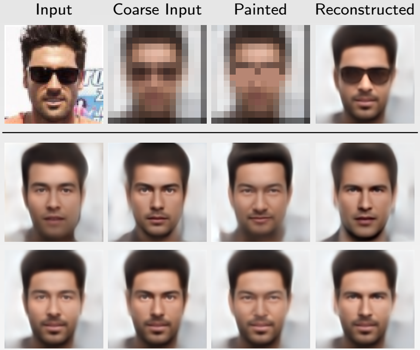
Editing: removing glasses

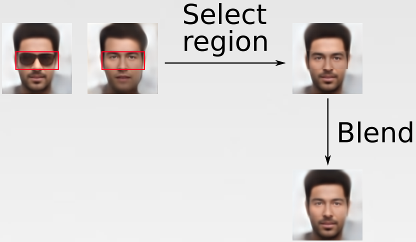
Input | Coarse Input | Painted | Reconstructed



Blend samples

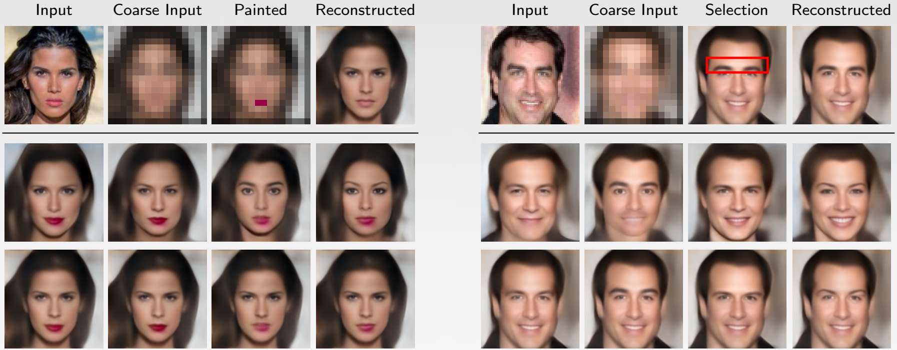
Editing: removing glasses





Blend samples

Editing: more examples



Add lipstick and adjust eyebrows

Conclusions

Conclusions

- Presented a conditional multi-scale extension of VAE
- Reconstructions and samples are sharper than VAE
- Model allows partial sampling

Limitations and extensions

- Greedy learning
  - End-to-end training strategies
- Gaussian likelihood
  - Complex distributions: perceptual loss or PixelCNN layers

Thank you

Questions?

References

[1] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[2] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[3] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

[5] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *ICML*, 2014.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[7] David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *CoRR*, 2017.

[8] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1558–1566. JMLR, 2016.

[9] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. PixelVAE: A Latent Variable Model for Natural Images. In *International Conference on Learning Representations (ICLR)*, 2017.