Diverse super-resolution with pretrained hierarchical VAEs

Jean Prost[1] Antoine Houdard[1] Andrés Almansa[2] Nicolas Papadakis[1]

Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251,F-33400 Talence, France

Université de Paris, MAP5, CNRS, F-75006 Paris, France

December 7, 2022

Outline

1. Variational autoencoders
2. Plugging the VAE prior in a restoration model
3. Application: Super-resolution with VDVAE
4. Results

Variational autoencoders

Variational autoencoders (VAEs)[1] allows to learn latent variable models of the form:

$$p_\theta(\boldsymbol{x}) = \int \underbrace{p_\theta(\boldsymbol{x}|\boldsymbol{z})}_{\text{decoder}} \underbrace{p_\theta(\boldsymbol{z})}_{\text{(latent) prior}} d\boldsymbol{z},$$

jointly with an approximate posterior distribution (encoder):

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) \approx p_\theta(\boldsymbol{z}|\boldsymbol{x})$$

VAE are trained to maximize the evidence lower-bound (ELBO):

$$\log p_\theta(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|x)}\right] - KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{x}|\boldsymbol{z})) \tag{1}$$

$$\geq \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|x)}\right]}_{\mathcal{L}_{\theta,\phi}(\boldsymbol{x})} \tag{2}$$

[1] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

Hierarchical VAE

Hierarchical prior

$$p_\theta(\boldsymbol{z}) = p_\theta(\boldsymbol{z}_0) \prod_{l=1}^{L-1} p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l}) \tag{3}$$

with $\boldsymbol{z}_l \in \mathbb{R}^{C_l \times H_l \times W_l}$, and

$$p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l}) \sim \mathcal{N}(\boldsymbol{z}_l|\mu_\theta(\boldsymbol{z}_l), \Sigma_\theta(\boldsymbol{z}_l)) \tag{4}$$

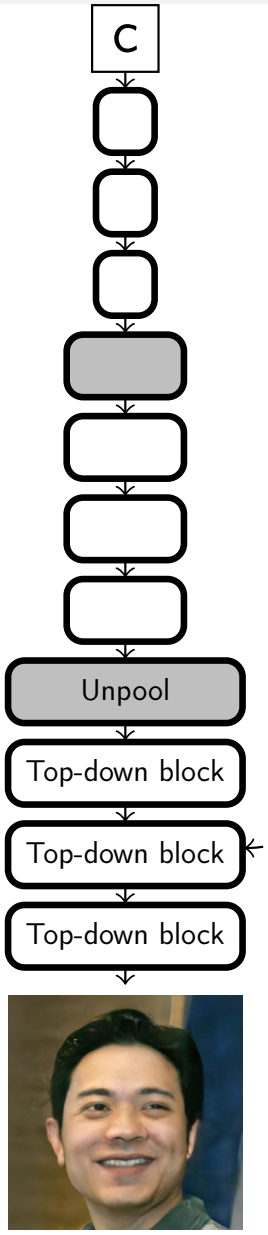
Hierarchical encoder[2]:

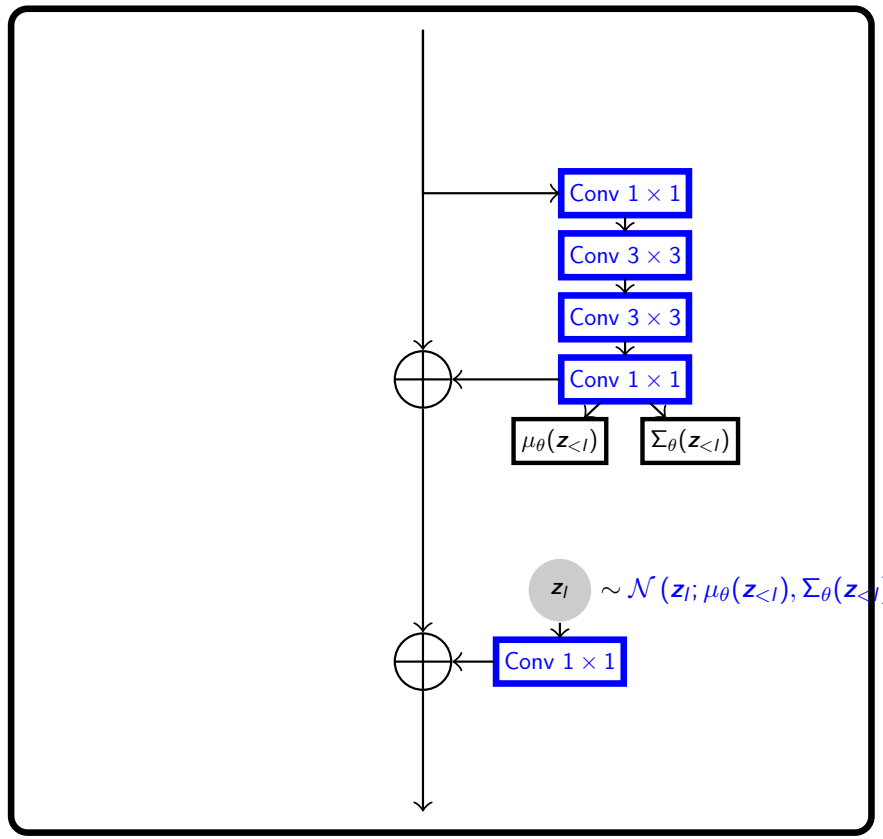
$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = q_\psi(\boldsymbol{z}_0|\boldsymbol{x}) \prod_{l=1}^{L-1} q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{<l}, \boldsymbol{x}) \tag{5}$$

[2] Casper Kaae Sø nderby et al. "How to train deep variational autoencoders and probabilistic ladder networks". In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016.

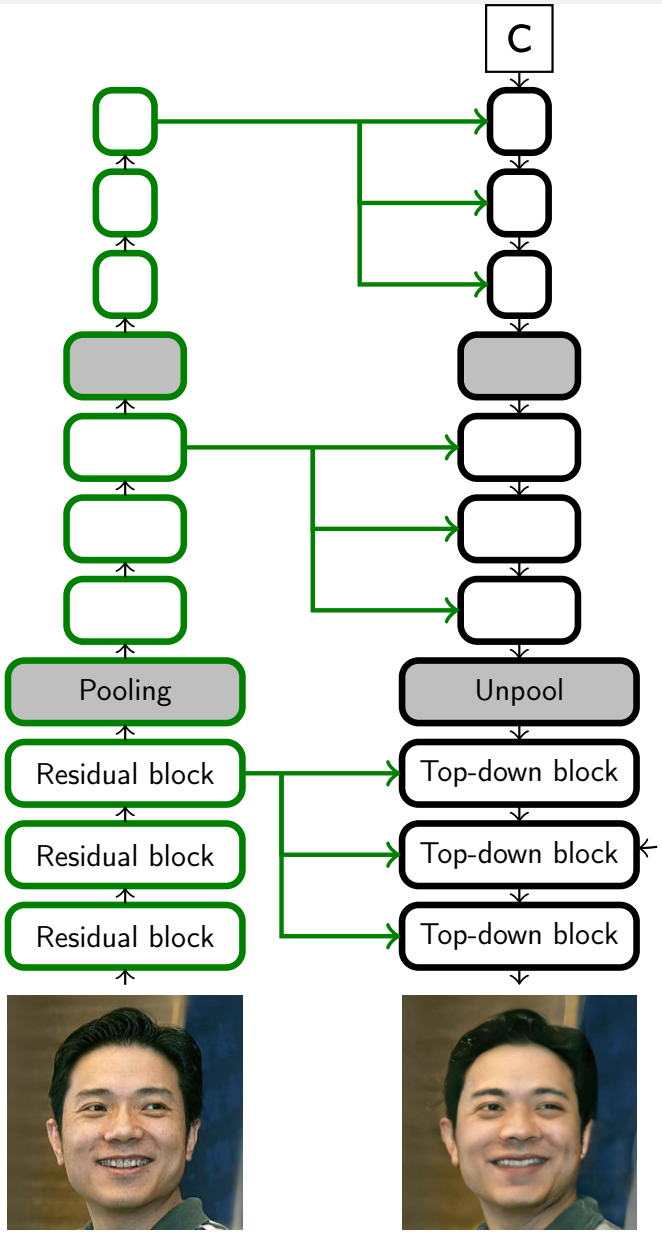
Hierarchical VAE: Architecture

$$p_\theta(z, x) = p_\theta(z_0) \prod_{l=1}^{L-1} p_\theta(z_l | z_{<l}) p_\theta(x|z)$$



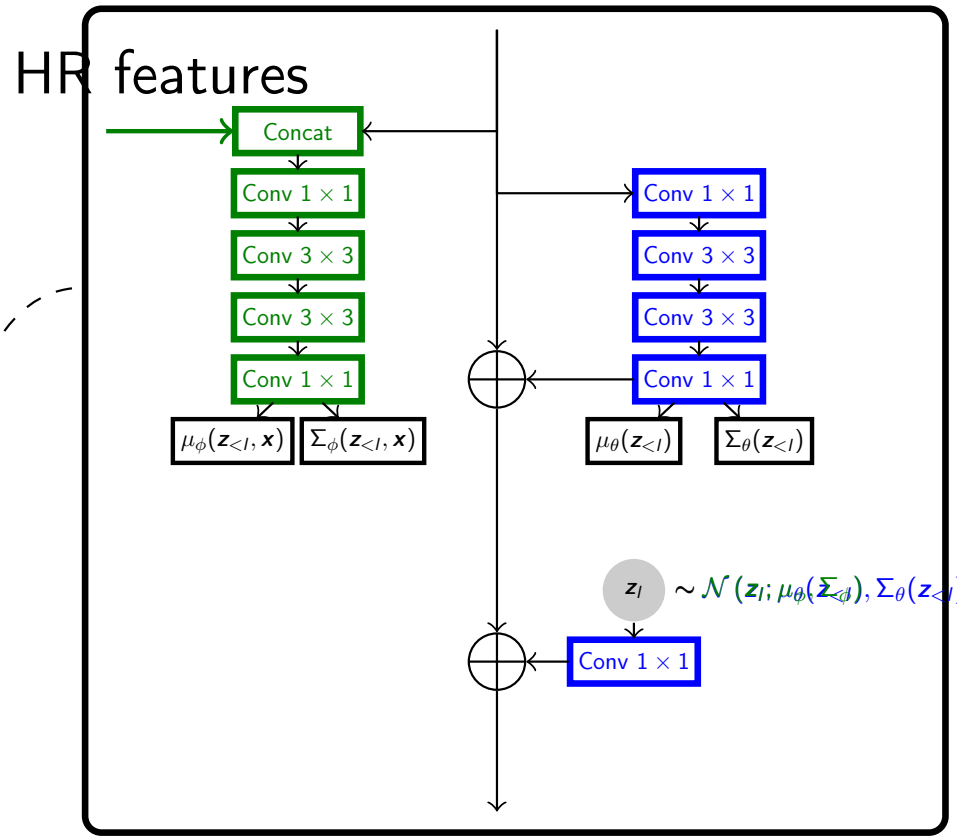


Hierarchical VAE: Architecture



$$p_\theta(z, x) = p_\theta(z_0) \prod_{l=1}^{L-1} p_\theta(z_l|z_{<l}) p_\theta(x|z)$$

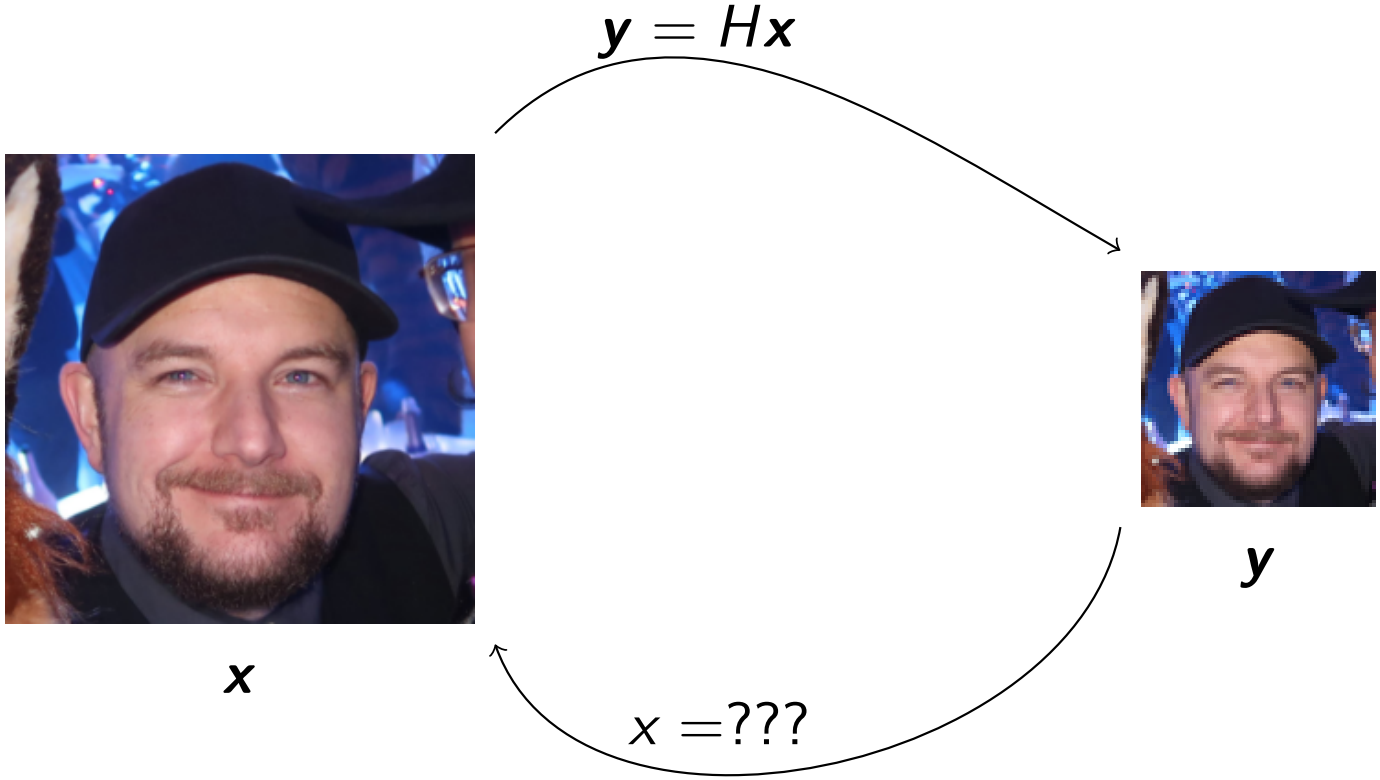
$$q_\phi(z|x) = q_\phi(z_0|x) \prod_{l=1}^{L-1} q_\phi(z_l|z_{<l}, x)$$



Outline

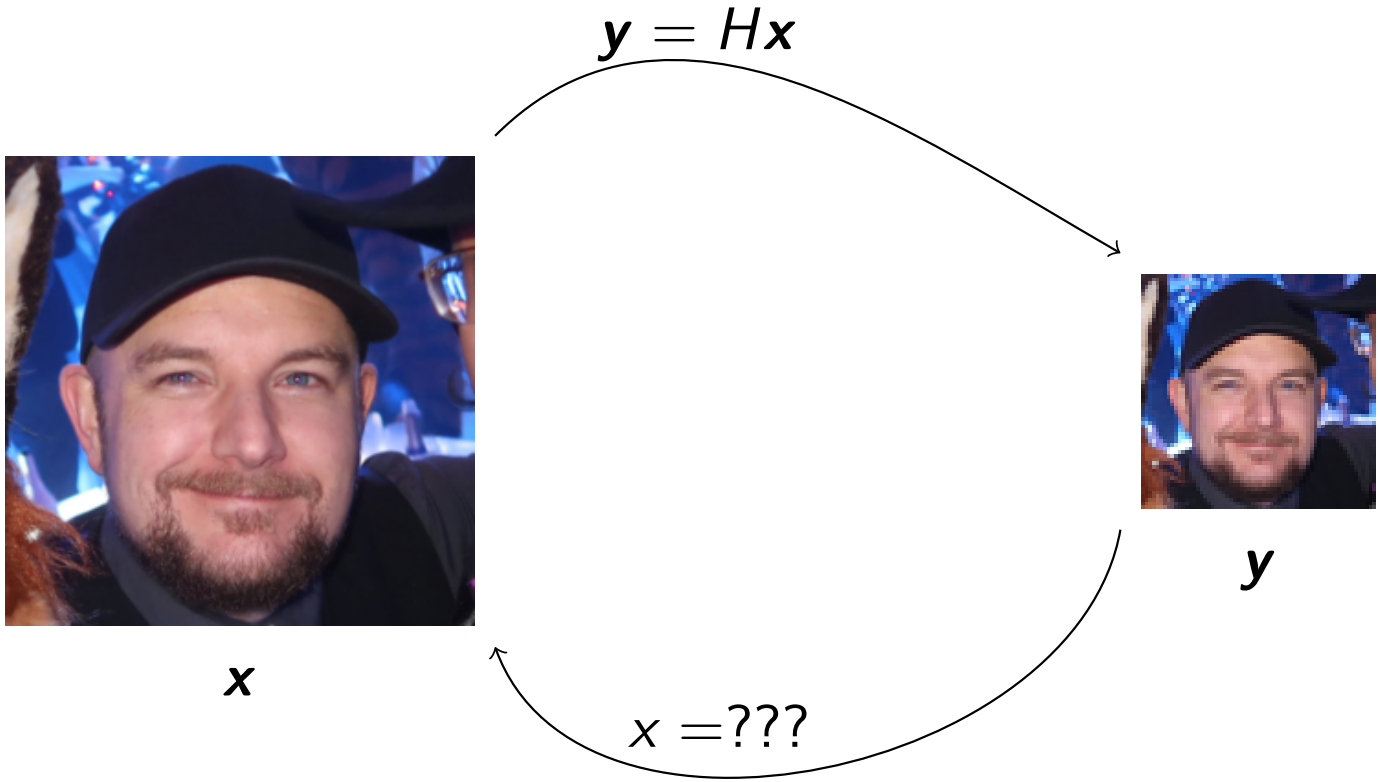
1. Variational autoencoders
2. Plugging the VAE prior in a restoration model
3. Application: Super-resolution with VDVAE
4. Results

Super-resolution



- High-resolution (HR) image ($\boldsymbol{x}$)
- Low-resolution (LR) image ($\boldsymbol{y}$)
- Forward operator H (low-pass + subsampling)

Super-resolution



- One-to-many problem : there can be many good solutions (consistent and realistic)
- Goal : model the distributions of solutions : $p_{SR}(x|y)$

Plugging the VAE prior in a restoration model

Target distribution:

$$p_\theta(\boldsymbol{x}|\boldsymbol{y}) :\propto p(\boldsymbol{y}|\boldsymbol{x}) \underbrace{p_\theta(\boldsymbol{x})}_{\text{VAE prior}} \tag{6}$$

$$p_\theta(\boldsymbol{x}|\boldsymbol{y}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y}) p_\theta(\boldsymbol{z}|\boldsymbol{y}) d\boldsymbol{z} \tag{7}$$

$$= \int \underbrace{p_\theta(\boldsymbol{x}|\boldsymbol{z})}_{\text{VAE decoder}} p_\theta(\boldsymbol{z}|\boldsymbol{y}) d\boldsymbol{z} \tag{8}$$

$\rightarrow$ When $p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{y}) = p_\theta(\boldsymbol{x}|\boldsymbol{z})$, we only need to model $p_\theta(\boldsymbol{x}|\boldsymbol{y})$

Learning the low-resolution encoder

Model distribution :

$$p_{SR}(\boldsymbol{x}|\boldsymbol{y}) = \int \underbrace{p_\theta(\boldsymbol{x}|\boldsymbol{z})}_{\text{decoder}} \underbrace{q_\psi(\boldsymbol{z}|\boldsymbol{y})}_{\text{LR encoder}} \, d\boldsymbol{z} \tag{9}$$

Conditional log likelihood of the model[3]:

$$\log p_{SR}(\boldsymbol{x}|\boldsymbol{y}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||q_\psi(\boldsymbol{z}|\boldsymbol{y})) \tag{10}$$

$\rightarrow$ Loss function:

$$\mathbb{E}_{p_\mathcal{D}(\boldsymbol{x},\boldsymbol{y})}\left[KL\left(\underbrace{q_\phi(\boldsymbol{z}|\boldsymbol{x})}_{\text{HR encoder}} \;||\; \underbrace{q_\psi(\boldsymbol{z}|\boldsymbol{y})}_{\text{LR encoder}} \right) \right] \tag{11}$$

[3] William Harvey, Saeid Naderiparizi, and Frank Wood. "Conditional Image Generation by Conditioning Variational Auto-Encoders". In: *International Conference on Learning Representations*. 2022.

Outline

1. Variational autoencoders
2. Plugging the VAE prior in a restoration model
3. Application: Super-resolution with VDVAE
4. Results

For deep HVAE, the low-frequency information is only encoded at the beginning of the hierarchy, in the low-resolution latents:

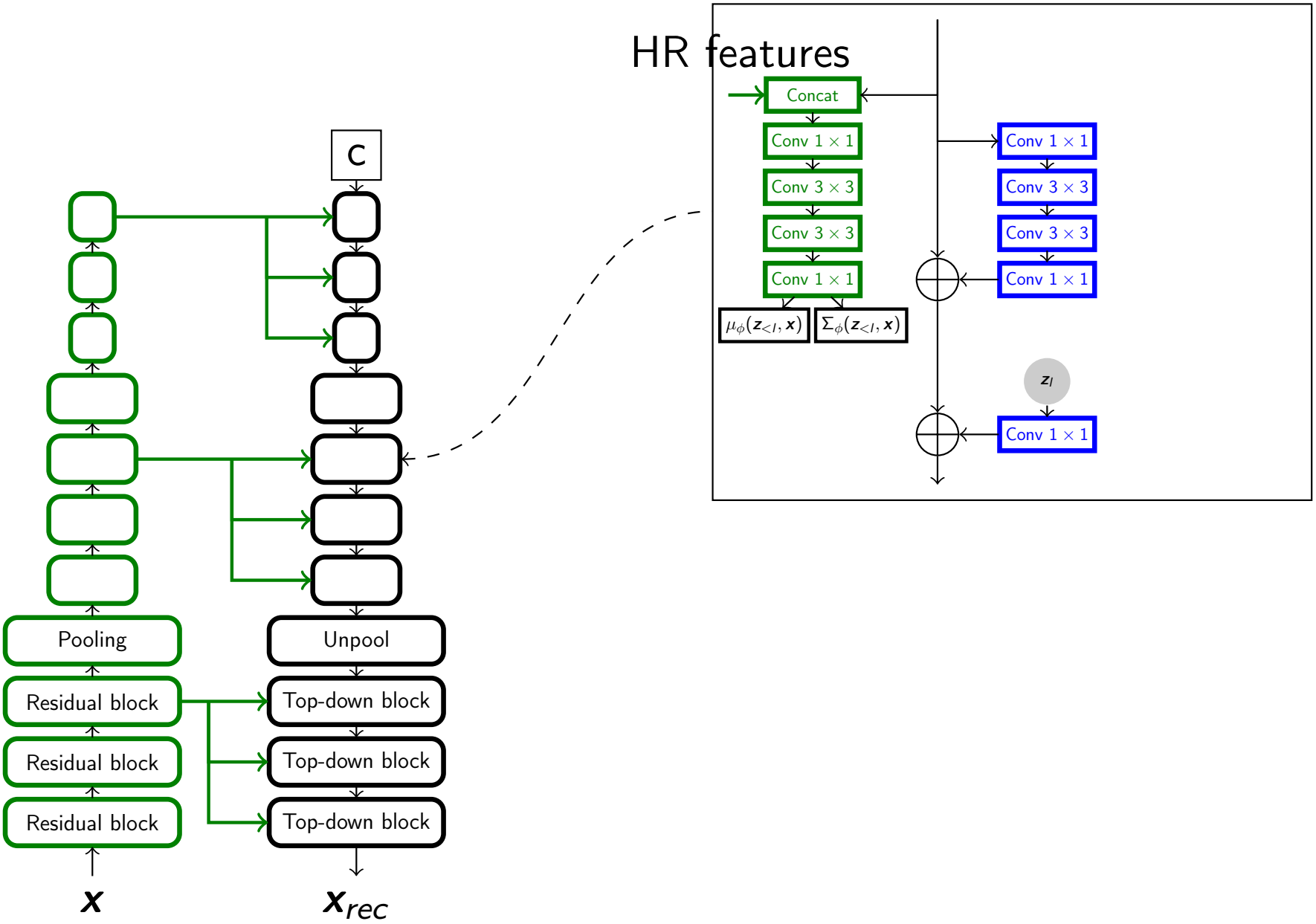
Figure: Samples from VDVAE[4] generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<43})$. Right: pixel-wise standard deviation.

$\rightarrow$ $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<k}, y) = p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<k})$.

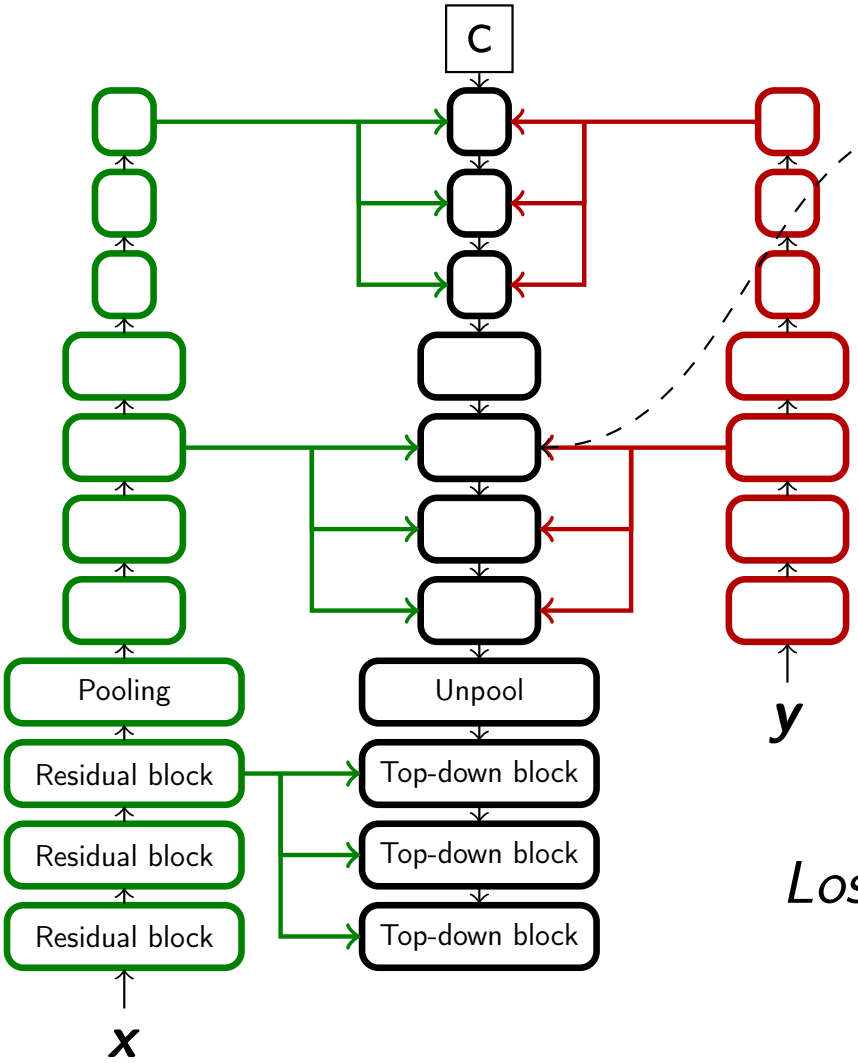
$\rightarrow$ We just have to learn $q_\psi(\boldsymbol{z}_{<k}|\boldsymbol{y})$

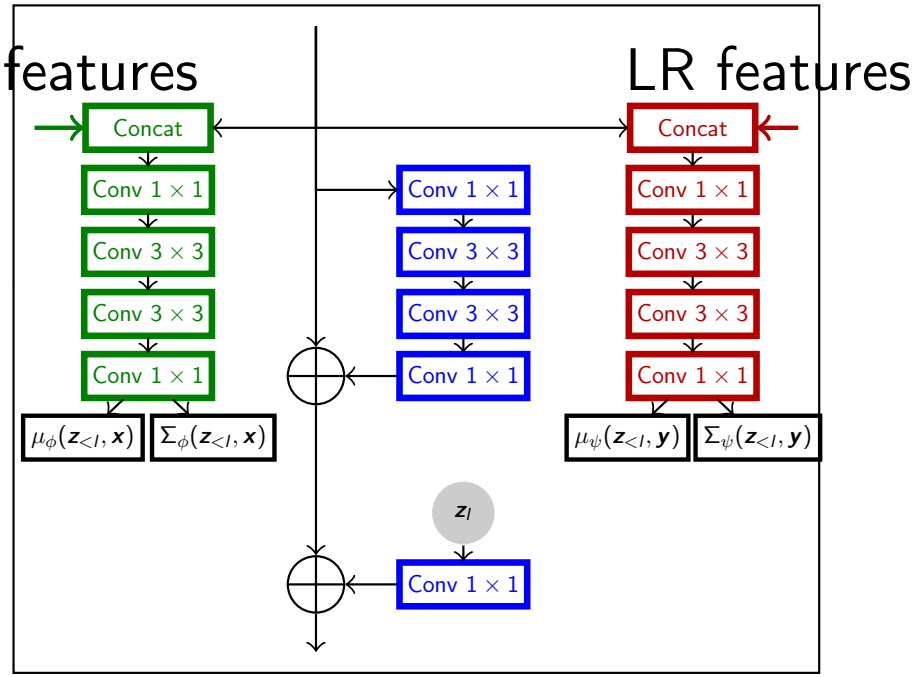
[4] Rewon Child. "Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images". In: *arXiv preprint arXiv:12011.10650* (2021).

Low-resolution encoder (super-resolution)



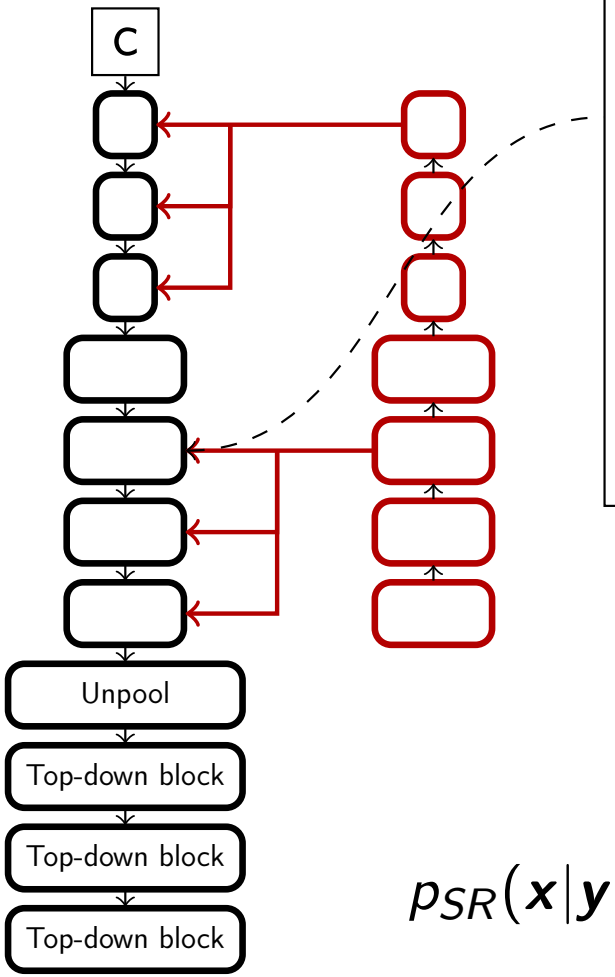
Low-resolution encoder (super-resolution)

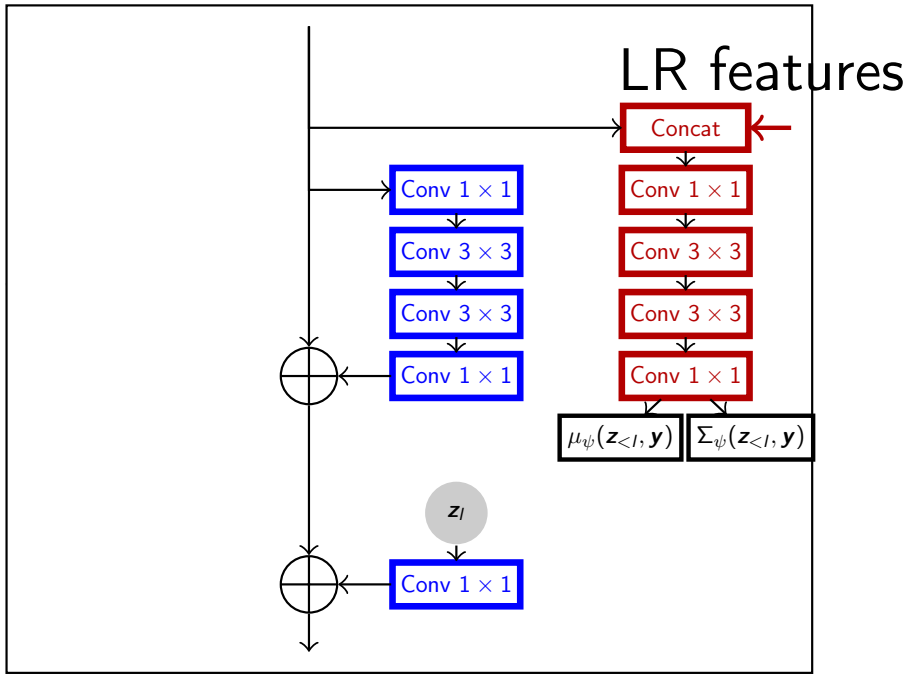




$$Loss(\boldsymbol{x}, \boldsymbol{y}, \psi) = KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||q_\psi(\boldsymbol{z}|\boldsymbol{y}))$$

Low-resolution encoder (super-resolution)





$$p_{SR}(x|y) = \int p_\theta(x|z_{<k}) q_\psi(z_{<k}|x) dz_{<k}$$

Outline

1. Variational autoencoders
2. Plugging the VAE prior in a restoration model
3. Application: Super-resolution with VDVAE
4. Results

Experiments

- Super-resolution on FFHQ256 dataset
- $\times 4$, $\times 8$, $\times 16$ upscaling
- Comparaison with HCFlow[5] and bicubic interpolation

[5] Jingyun Liang et al. "Hierarchical Conditional Flow: A Unified Framework for Image Super-Resolution and Image Rescaling". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 4076–4085.

Results

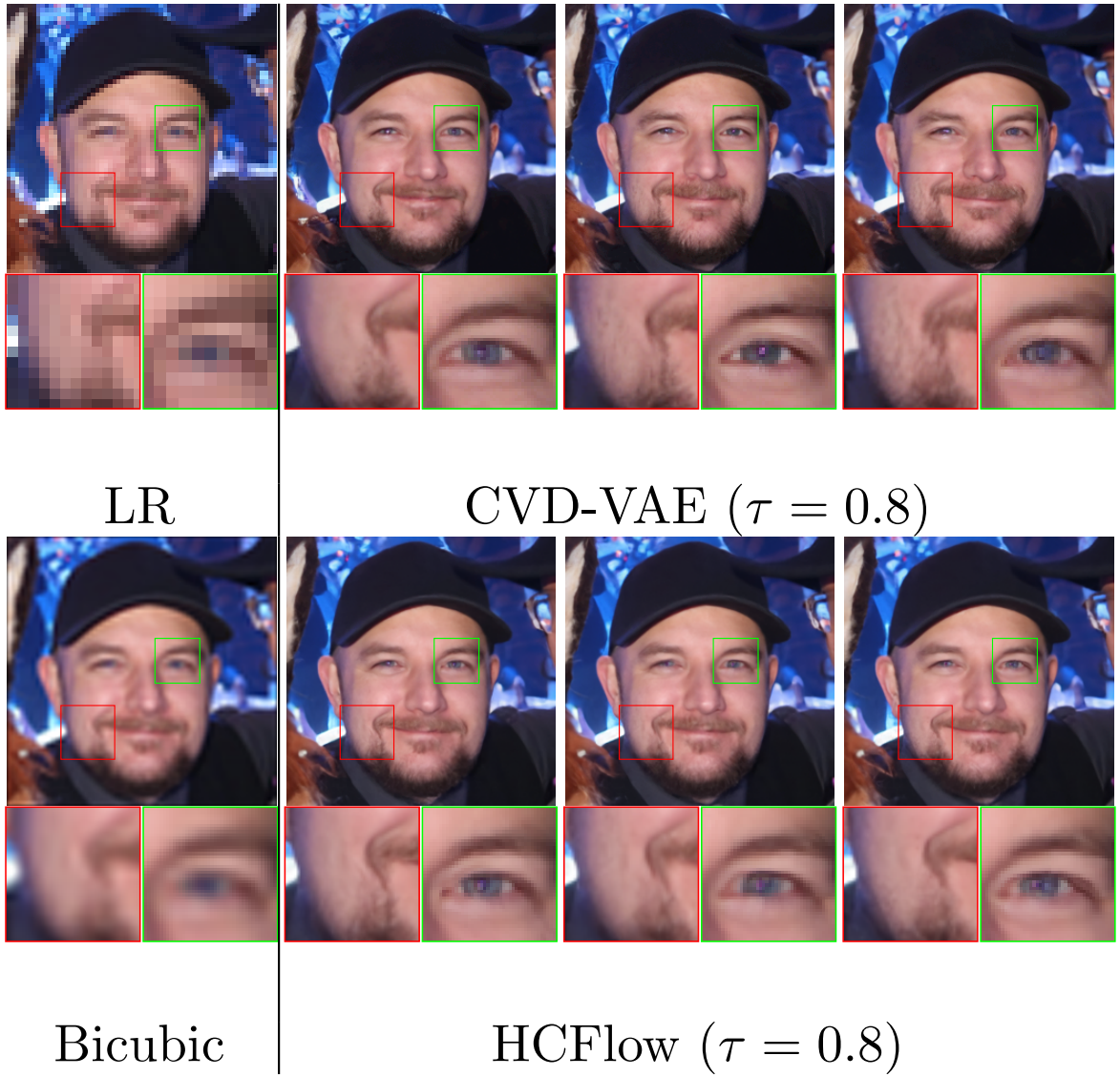


Figure: $\times 4$

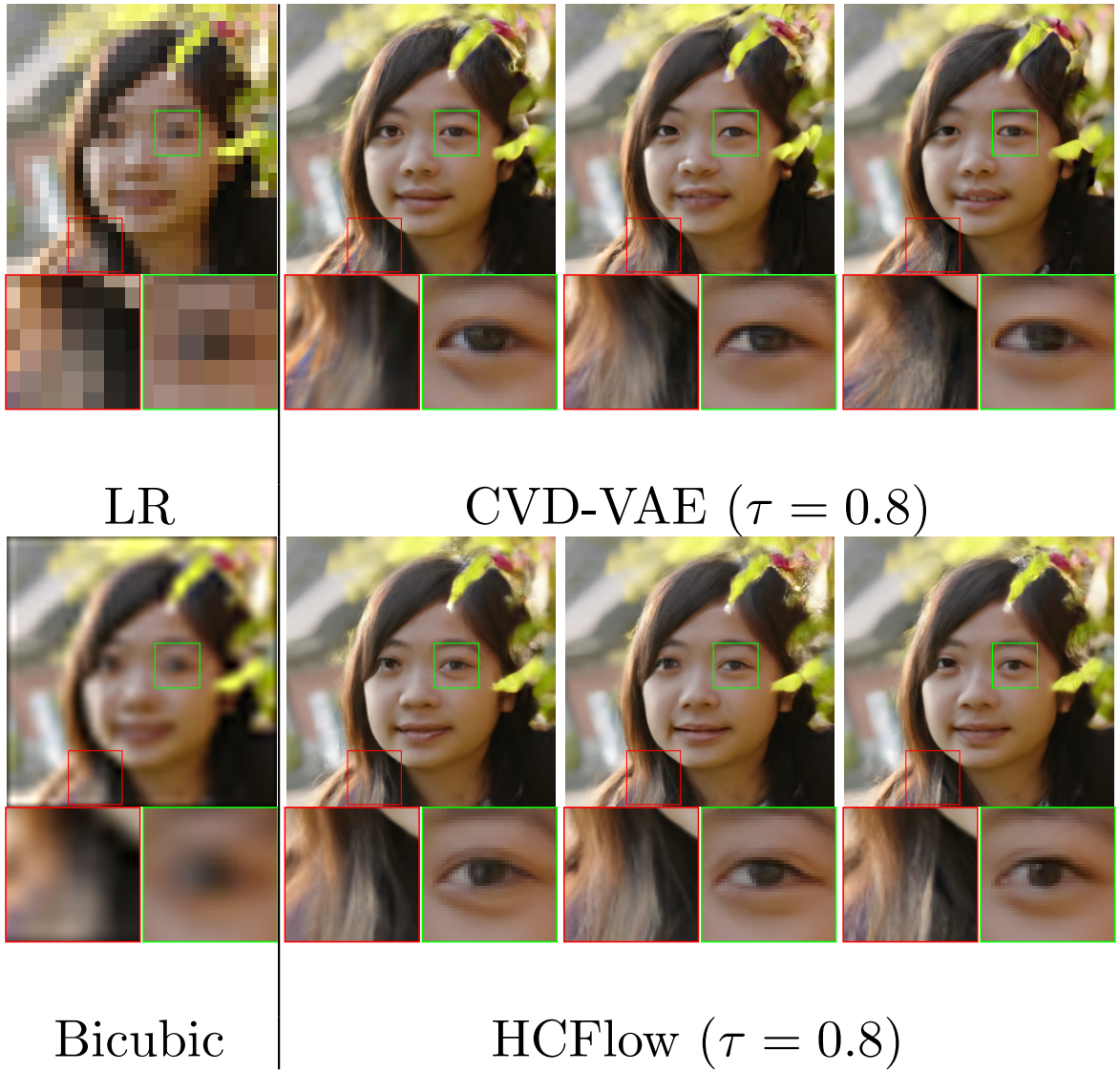


Figure: $\times 8$

Results

		Visual Quality	Consistency	Diversity	
		BRISQUE↓	LR-PSNR ↑	APD (MSE) ($\times 10^4$) ↑	APD (LPIPS) ($\times 10^3$) ↑
×4	Bicubic	61.79	36.99	0	0
	HCFlow ($\tau = 0.1$)	48.21	52.66	1.3	1.1
	HCFlow ($\tau = 0.8$)	37.21	52.81	**161.8**	62.6
	CVD-VAE ($\tau = 0.1$)	<u>36.47</u>	**75.70**	64.6	<u>104.5</u>
	CVD-VAE ($\tau = 0.8$)	**32.30**	<u>75.20</u>	<u>88.8</u>	**123.0**
×8	Bicubic	78.42	33.61		
	HCFlow ($\tau = 0.1$)	69.05	51.35	4.9	4.2
	HCFlow ($\tau = 0.8$)	<u>36.25</u>	51.13	**575.5**	155.3
	CVD-VAE ($\tau = 0.1$)	50.34	**71.63**	140.4	<u>179.0</u>
	CVD-VAE ($\tau = 0.8$)	**32.26**	<u>70.15</u>	<u>248.2</u>	**236.4**
×16	Bicubic	97.53	30.67		
	HCFlow ($\tau = 0.1$)	74.28	51.93	3.3	14.6
	HCFlow ($\tau = 0.8$)	**30.83**	53.81	<u>323.7</u>	<u>268.6</u>
	CVD-VAE ($\tau = 0.1$)	52.75	**60.68**	299.4	242.4
	CVD-VAE ($\tau = 0.8$)	<u>32.01</u>	<u>57.02</u>	**613.8**	**341.2**

APD = Average Pairwise Distance

Effect of the temperature

Figure: Effect of the sampling temperature τ

Figure: Reducing the temperature can help to reduce artifacts for difficult images.

More samples

Figure: $\times 4$

More samples

Figure: $\times 8$

Conclusion

- Hierarchical VAE can model complex image priors that can be used to solve inverse problems.
- For future works:
  - Apply on different datasets
  - Flexible method (for general inverse problems)

Expected consistency of the super-resolution

Expected consistency of $p_{SR}(\boldsymbol{x}|\boldsymbol{y})$

$$CE(k) = \mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{y})}\mathbb{E}_{p_{SR}(\boldsymbol{x}|\boldsymbol{y})}\left[\frac{1}{\sqrt{m}}\|H\boldsymbol{x} - \boldsymbol{y}\|_2\right]. \quad (12)$$

Proposition

If :

1. The low-resolution encoder has enough capacity and is trained to optimality
2. The VAE encoder $q_\phi(\boldsymbol{x}|\boldsymbol{z})$ and generative model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$ have enough capacity and are trained to optimality

Then:

$$CE(k) = \mathbb{E}_{p_\theta(\boldsymbol{z}_{<k})}\mathbb{E}_{p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<k})}\mathbb{E}_{p_\theta(\tilde{\boldsymbol{x}}|\boldsymbol{z}_{<k})}\left[\frac{1}{\sqrt{m}}\|H\tilde{\boldsymbol{x}} - H\boldsymbol{x}\|_2\right] := U_k^s. \quad (13)$$

Expected consistency of the super-resolution model

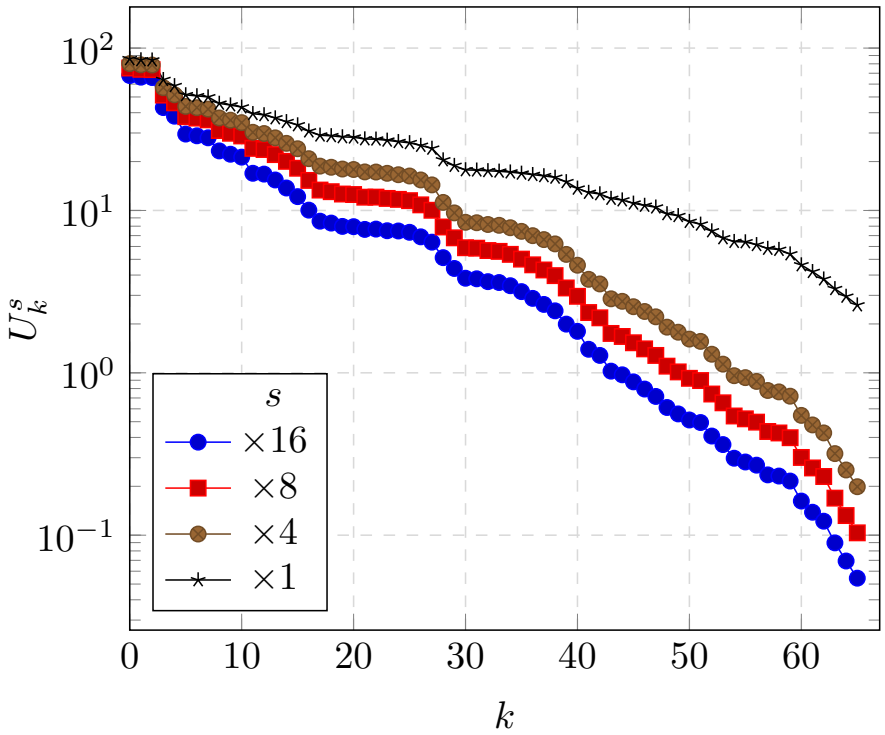


Figure: Average low-resolution pairwise distance, U_k^s between samples from the conditional generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<k})$ of VD-VAE, for downscaling factors $s = 1, 4, 8, 16$. Image with pixel values in $[0, 255]$.

Enforcing the consistency with the input

Projection to space of consistent solution $\{\boldsymbol{x}|H\boldsymbol{x} = \boldsymbol{y}\}$:

$$\hat{\boldsymbol{x}} \leftarrow (I - H^T(HH^T)^{-1}H)\boldsymbol{x} + H^T(HH^T)^{-1}\boldsymbol{y} \tag{14}$$

We use the CEM implementation of[6]

[6] Yuval Bahat and Tomer Michaeli. "Explorable super resolution". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 2716–2725.

Effect of the projection

		Distortion			Visual Quality	Consistency	Diversity	
	model	PSNR↑	SSIM↑	LPIPS ↓	BRISQUE↓	LR-PSNR ↑	APD (MSE) ($\times 10^4$) ↑	APD (LPIPS) ($\times 10^3$)
$\times 4$	CVD-VAE ($\tau = 0.1$)	29.09	0.839	0.177	34.56	35.31	77.5	123.6
	CVD-VAE ($\tau = 0.8$)	28.66	0.825	0.177	**30.37**	35.09	**105.7**	**142.9**
	CVD-VAE ($\tau = 0.1$) (cons)	**30.75**	**0.863**	**0.152**	36.47	**75.70**	64.6	104.5
	CVD-VAE ($\tau = 0.8$) (cons)	30.24	0.850	0.157	32.30	75.20	88.8	123.0
$\times 8$	CVD-VAE ($\tau = 0.1$)	25.46	0.735	0.312	47.48	33.93	181.6	200.7
	CVD-VAE ($\tau = 0.8$)	24.71	0.697	0.284	**30.40**	33.42	**308.0**	**256.2**
	CVD-VAE ($\tau = 0.1$) (cons)	**26.27**	**0.747**	0.299	50.34	**71.63**	140.4	179.0
	CVD-VAE ($\tau = 0.8$) (cons)	25.47	0.708	**0.275**	32.26	70.15	248.2	236.4
$\times 16$	CVD-VAE ($\tau = 0.1$)	20.16	0.601	0.440	47.34	25.20	711.1	283.1
	CVD-VAE ($\tau = 0.8$)	19.21	0.541	0.401	**29.17**	24.22	**1278.8**	**374.0**
	CVD-VAE ($\tau = 0.1$) (cons)	**22.06**	**0.622**	0.427	52.75	**60.68**	299.4	242.4
	CVD-VAE ($\tau = 0.8$) (cons)	21.14	0.564	**0.388**	32.01	57.02	613.8	341.2

Figure: Effect of the projection

Effect of the projection

$\times 4$ $\times 8$ $\times 16$

From top row to bottom row : LR, SR, SR projected

LR SR

SR (projected)

Figure: $\times 16$

More samples ($\times 16$)

Figure: $\times 16$