Efficient Posterior Sampling for Diverse Super-Resolution with Hierarchical VAE Prior

Jean Prost[1], Antoine Houdard[2], Andrés Almansa[3] and Nicolas Papadakis[4]
[1]*Univ. Bordeaux, Bordeaux IMB, INP, CNRS, UMR 5251, F-33400 Talence, France*
[2]*Ubisoft La Forge, F-33000 Bordeaux, France*
[3]*Université Paris Cité, CNRS, MAP5, F-75006 Paris, France*
[4]*Univ. Bordeaux, CNRS, INRIA, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France*

Keywords: Diverse Image Super-Resolution, Hierarchical Variational Autoencoder, Conditional Generative Model.

Abstract: We investigate the problem of producing diverse solutions to an image super-resolution problem. From a probabilistic perspective, this can be done by sampling from the posterior distribution of an inverse problem, which requires the definition of a prior distribution on the high-resolution images. In this work, we propose to use a pretrained hierarchical variational autoencoder (HVAE) as a prior. We train a lightweight stochastic encoder to encode low-resolution images in the latent space of a pretrained HVAE. At inference, we combine the low-resolution encoder and the pretrained generative model to super-resolve an image. We demonstrate on the task of face super-resolution that our method provides an advantageous trade-off between the computational efficiency of conditional normalizing flows techniques and the sample quality of diffusion based methods.

1 INTRODUCTION

Image super-resolution is the task of generating a high-resolution (HR) image $\boldsymbol{x}$ corresponding to a low-resolution (LR) observation $\boldsymbol{y}$. A typical approach for image super-resolution is to train a deep neural network in a supervised fashion to map a LR image to its HR counterpart (see (Lepcha et al., 2022) for an extensive review). Despite impressive performances, those regression based methods are fundamentally limited by their lack of diversity. Indeed, there might exist many plausible HR solutions associated with one LR observation, but regression based methods only provide one of those solutions.

An alternative approach for image super-resolution is to sample from the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$. Specifically, we can train conditional deep generative models to fit the posterior $p(\boldsymbol{x}|\boldsymbol{y})$. With the recent advances in deep generative modeling, it is possible to generate realistic and diverse samples from the posterior distribution.

Starting from the seminal work of (Lugmayr et al., 2020), many approaches proposed to train conditional generative models such as conditional normalizing flow or conditional variational autoencoders in order to model the posterior distribution of the super-resolution problem. We refer to those methods as **direct methods**, as they only require one network function evaluation (NFE) to generate one sample.

With the recent development of score-based generative models (also known as denoising diffusion models) (Ho et al., 2020; Song et al., 2021), posterior sampling methods based on conditional denoising diffusion models are now able to produce high-quality samples outperforming previous direct methods (Choi et al., 2021; Chung et al., 2022; Kawar et al., 2022). However, denoising diffusion methods are limited by their computationally expensive sampling process, as they require numerous network function evaluations to generate one super-resolved sample. In the following, we classify those methods as iterative methods.

In this work, we address the question: Can we get the best of both worlds between the sampling quality of iterative methods, and the computationnal efficiency of direct methods? We show that it is indeed possible to reach this goal with our diverse super-resolution method CVDVAE (Conditional VDVAE).

Our approach is based on reusing a pretrained hierarchical variational autoencoder (HVAE) (Sønderby et al., 2016; Kingma et al., 2016). HVAE models are able to generate high-quality images by relying on an expressive sequential generative model (Vahdat and Kautz, 2020; Child, 2021; Hazami et al., 2022). By associating one latent variable subgroup to each residual block of a generative network, HVAE models are able to learn compact high-level representations of the data, and they can generate new samples efficiently, with only one evaluation of the generative network.

Prost, J., Houdard, A., Almansa, A. and Papadakis, N.
Efficient Posterior Sampling for Diverse Super-Resolution with Hierarchical VAE Prior.
DOI: 10.5220/0012352800003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 393-400
ISBN: 978-989-758-679-8; ISSN: 2184-4321

Figure 1: Diverse super-resolved samples at different upscaling factors. Our method can generate realistic samples with diverse attributes (hairs, mouth, eyes. . .).

The fast sampling time and the expressivity of HVAE models make them suitable candidates for efficient posterior sampling. In this work, we exploit a pretrained VD-VAE model (Child, 2021). In order to repurpose VD-VAE generative model for image super-resolution, we train a low-resolution encoder to encode LR images in the latent space of the VD-VAE model. By combining the LR encoder with the VD-VAE generative model, we can produce a sample with only one (autoencoder) network evaluation. By adopting a stochastic model for the LR encoder, our method can generate diverse samples from the posterior distribution, as illustrated in Figure 1. We show that the LR encoder can be trained with reasonable computational resources by exploiting the VD-VAE original (HR) encoder to generate labels for training the LR-encoder, and by sharing weights between the LR encoder and VD-VAE generative model. We evaluate our method on super-resolution of face images, with upscaling factor $\times 4$ and $\times 8$ and demonstrate that it reaches sample quality on par with sequential methods, while being significantly faster ($> \times 500$).

The paper is organised as follows. In section 2, we provide the necessary background on HVAE models. Then we present in section 3 our super-resolution method. Experimental results are given in section 4 and we discuss related works in section 5.

2 HIERARCHICAL VAE

Variational Autoencoder. We propose to use a hierarchical variational autoencoder as a prior model over high-resolution images. A variational autoencoder is a deep latent variable model of the form:

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{z}) p_\theta(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z}. \tag{1}$$

where $p_\theta(\boldsymbol{z})$ defines the prior distribution of the latent variable $\boldsymbol{z}$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is the decoding distribution.

A VAE also provides an inference model (encoder) $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, trained to match the intractable model posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ (Kingma and Welling, 2013). In order to define expressive models, both the generative model $p_\theta(\boldsymbol{z},\boldsymbol{x})$ and the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ are parameterized by neural networks, whose weights are respectively parameterized with θ and ϕ.

Hierarchical Generative Model. A hierarchical VAE is a specific class of VAE where the latent variable $\boldsymbol{z}$ is partitioned into L subgroups $\boldsymbol{z} = (\boldsymbol{z}_0, \boldsymbol{z}_1, \cdots, \boldsymbol{z}_{L-1})$, and the prior is set to have a hierarchical structure:

$$p_\theta(\boldsymbol{z}) = p_\theta(\boldsymbol{z}_0, \boldsymbol{z}_1, \cdots, \boldsymbol{z}_{L-1}) \tag{2}$$

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

In practice, each latent supgroup is a 3-dimensional tensor $\boldsymbol{z}_l \in \mathcal{R}^{c_l \times h_l \times w_l}$, with increasing resolution $h_0 \leq h_1 \leq \cdot \leq h_{L-1}$. Each conditional model in the hierarchical prior is set as a Gaussian:

$$\begin{cases} p_\theta(\boldsymbol{z}_0) & = \mathcal{N}(\boldsymbol{z}_0; \mu_{\theta,0}, \Sigma_{\theta,0}) \\ p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l}) & = \mathcal{N}(\boldsymbol{z}_l; \mu_{\theta,l}(\boldsymbol{z}_{<l}), \Sigma_{\theta,l}(\boldsymbol{z}_{<l})). \end{cases} \tag{4}$$

As illustrated in Figure 2, the generative model is embedded within a ”top-down” generative network. To generate an image, a low-resolution constant input tensor is sequentially processed by a serie of top-down blocks and upsampling layers (Figure 2a). In each top-down block l (Figure 2b), a latent subgroup $\boldsymbol{z}_l$ is sampled according to the statistics $\mu_{\theta,l}(\boldsymbol{z}_{<l})$ and $\Sigma_{\theta,l}(\boldsymbol{z}_{<l})$ computed within the top-down block.

Hierarchical Encoder. The HVAE encoder has the same hierarchical structure as the generative model:

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

with Gaussian parameterization of the conditional distributions:

$$\begin{cases} q_\phi(z_0|\boldsymbol{x}) &= \mathcal{N}\left(z_0;\mu_{\phi,0}(\boldsymbol{x}),\Sigma_{\phi,0}(\boldsymbol{x})\right) \\ q_\phi(z_l|z_{<l},\boldsymbol{x}) &= \mathcal{N}\left(z_l;\mu_{\phi,l}\left(z_{<l},\boldsymbol{x}\right),\Sigma_{\phi,l}(z_{<l},\boldsymbol{x})\right). \end{cases} \tag{6}$$

The HVAE encoder is merged with the generative top-down network. In each top-down block l, a branch associated with the encoder uses features from the input $\boldsymbol{x}$ along with features from the previous levels to infer the encoder statistics at level l (in green in Figure 2b). The image features are extracted by a "bottom-up" network (in green in Figure 2a).

3 SUPER-RESOLUTION HVAE

Problem Formulation. In this section we describe our super-resolution method based on a pretrained HVAE model. We assume that the LR image $\boldsymbol{y} \in \mathbb{R}^{3\times\frac{H}{s}\times\frac{W}{s}}$, and its associated HR image $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$ are related by a linear degradation model:

$$\boldsymbol{y} = (k * \boldsymbol{x}) \downarrow_s, \tag{7}$$

where k is a low-pass filter and $(u)\downarrow_s$ is defined as the subsampling operation with downsampling factor s. Our goal is to sample from the posterior distribution of the inverse problem:

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}). \tag{8}$$

In (8), the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ can be deduced from the degradation model (7). On the other hand, the prior model $p(\boldsymbol{x})$ needs to be specified by the user. Deep generative models such as GANs, VAE or diffusion models can be used to model the prior on high-resolution images. In the following, we propose to parameterize $p(\boldsymbol{x})$ with a hierarchical variational autoencoder. Given a pretrained HVAE prior $p_\theta(\boldsymbol{x})$, the ideal super-resolution model is:

$$p_\theta(\boldsymbol{x}|\boldsymbol{y}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})p_\theta(\boldsymbol{z}|\boldsymbol{y})d\boldsymbol{z}, \tag{9}$$

where probability laws correspond to the conditional of the augmented model $p_\theta(\boldsymbol{z},\boldsymbol{x},\boldsymbol{y}) := p_\theta(\boldsymbol{z})p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{x})$. Since we do not have acces to $p_\theta(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})$ and $p_\theta(\boldsymbol{z}|\boldsymbol{y})$, we can not directly sample from (9). However, we will see in the following part that we can efficiently approximate this model by making use of the structure of the HVAE hierarchical latent representation and of the its pretrained encoder.

Super-resolution Model. It has been observed in several works that the low-frequency information of images generated by HVAE model where mostly controlled by the low-resolution latent variable, at the beginning of the hierarchy (Vahdat and Kautz, 2020; Child, 2021; Havtorn et al., 2021). Hence, for a large enough number of latent groups k, samples from $p_\theta(\boldsymbol{x}|\boldsymbol{x}_{<k})$ share the same low-frequency information. As a consequence, all the samples from $p_\theta(\boldsymbol{x}|\boldsymbol{x}_{<k})$ are consistent to a LR image $\boldsymbol{y}$ (up to a small error).

This motivates us to define the following super-resolution model:

$$p_{SR}(\boldsymbol{x}|\boldsymbol{y}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<k})q_\psi(\boldsymbol{z}_{<k}|\boldsymbol{y})d\boldsymbol{z}, \tag{10}$$

where $q_\psi(\boldsymbol{z}|\boldsymbol{y})$ is a stochastic low-resolution encoder, trained to encode the low-resolution latent groups. By definition of the super-resolution model (10), we can sample from $p_{SR}(\boldsymbol{x}|\boldsymbol{y})$ by sequentially sampling $\boldsymbol{z}_{<k} \sim q_\psi(\boldsymbol{z}_{<k}|\boldsymbol{y})$ and $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<k})$.

Hierarchical Low-Resolution Encoder. We set the LR encoder to have a hierarchical structure:

$$q_\psi(\boldsymbol{z}_{<k}|\boldsymbol{y}) = q_\psi(z_0|\boldsymbol{y})\prod_{l=1}^{k} q_\psi(z_l|\boldsymbol{z}_{<l},\boldsymbol{y}), \tag{11}$$

with Gaussian conditional distributions:

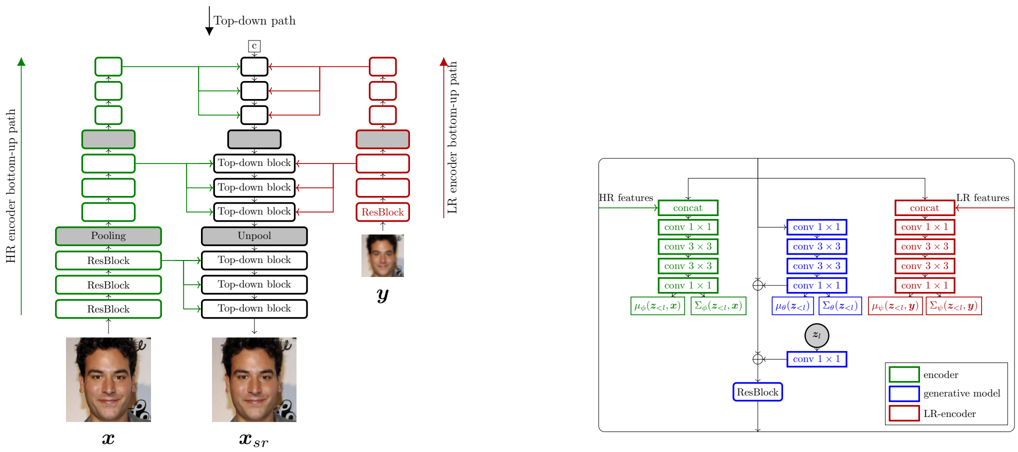
$$\begin{cases} q_\psi(z_0|\boldsymbol{y}) &= \mathcal{N}\left(z_0;\mu_{\psi,0}(\boldsymbol{y}),\Sigma_{\psi,0}(\boldsymbol{y})\right) \\ q_\psi(z_l|z_{<l},\boldsymbol{y}) &= \mathcal{N}\left(z_l;\mu_{\psi,l}(z_{<l},\boldsymbol{y}),\Sigma_{\psi,l}(z_{<l},\boldsymbol{y})\right). \end{cases} \tag{12}$$

We implement the LR encoder with the same architecture as VD-VAE original (HR) encoder, but with a limited number of blocks due to reduced number of latent variable to be predicted (Figure 2). Only the parameters of the low-resolution encoder (in red in Figure 2) are trained, while the shared parameters (in blue in Figure 2) are set to the value of the corresponding parameters in the pretrained VD-VAE generative model, and remain frozen during training.

Training. We keep the weights of the HVAE decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, so that the only trainable weights of our super-resolution model (10) are the weights of the LR encoder ψ. Given a joint training distribution of HR-LR image pairs $p_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{y})$, the LR encoder is trained to match the available "high-resolution" HVAE encoder $q_\phi(\boldsymbol{z}_{<k}|\boldsymbol{x})$ on the associated HR images, by minimizing the Kullback-Leibler (KL) divergence:

$$\mathcal{L}(\psi) = \mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{y})}\left[KL\left(q_\phi(\boldsymbol{z}_{<k}|\boldsymbol{x})||q_\psi(\boldsymbol{z}_{<k}|\boldsymbol{y})\right)\right]. \tag{13}$$

The criterion (13) was introduced by (Harvey et al., 2022), who demonstrated that minimizing (13) is equivalent to maximizing a lower-bound of the super-resolution conditional log-likelihood on the training dataset, and that under additional assumptions on the pretrained HVAE model, one can reach optimal performance by only training the low-resolution encoder $q_\psi(\boldsymbol{z}_{<k}|\boldsymbol{y})$. In practice, the KL divergence within the



(a) CVDVAE architecture (b) Top-down block

Figure 2: Super-resolution model based on a pretrained VD-VAE model. (a) Given a pretrained VD-VAE network, composed of an encoder bottom-up network (in green), and a top-down network (in black), we train a LR encoder (in red) to encode low-resolution images in the latent space of VD-VAE. The LR-encoder has the same structure than VD-VAE encoder, with a bottom-up path that extracts multiscale features, and a top-down path merged with VD-VAE top-down path. (b) In each block of the top-down path, we add branch for the LR encoder (in red) to infer the statistics of $q_\psi(\boldsymbol{z}_l|\boldsymbol{z}_{<l},\boldsymbol{y})$.

training criterion (13) can be decomposed into a sum of KL divergence on each latent subgroup:

$$KL(q_\phi(\boldsymbol{z}_{<k}|\boldsymbol{x})||q_\psi(\boldsymbol{z}_{<k}|\boldsymbol{y})) = KL(q_\phi(z_0|\boldsymbol{x})||q_\psi(z_0|\boldsymbol{y})) + \mathbb{E}_{q_\phi(\boldsymbol{z}_{<k}|\boldsymbol{x})}\left[\sum_{l=1}^{k} KL(q_\phi(z_l|\boldsymbol{z}_{<l},\boldsymbol{x})||q_\psi(z_l|\boldsymbol{z}_{<l},\boldsymbol{y}))\right]. \quad (14)$$

Since each conditional law involved in (14) is Gaussian, each KL term can be computed in closed-form. In practice the covariance matrices $\Sigma_{\phi,l}(\boldsymbol{z}_{<l},\boldsymbol{x})$ and $\Sigma_{\psi,l}(\boldsymbol{z}_{<l},\boldsymbol{y})$ are constrained to be diagonal, so that the KL can be computed efficiently.

4 EXPERIMENTS

4.1 Experimental Settings

Dataset and Upscaling Factors. We test our super-resolution method CVDVAE on the FFHQ dataset (Karras et al., 2019), with images of resolution 256×256. We experiment on 2 upscaling factors: $\times 4$ ($64 \times 64 \rightarrow 256 \times 256$) and $\times 8$ ($32 \times 32 \rightarrow 256 \times 256$). The low resolution images are initially downscaled by applying an antialiasing kernel followed by a bicubic interpolation.

Compared Methods. We compare CVDVAE with a conditional normalizing flow (HCFlow) (Liang et al., 2021), a conditional diffusion model (SR3) (Saharia et al., 2021), and a method that add guidance to a non-conditional diffusion model at inference (DPS) (Chung et al., 2022). We retrain HCFlow on FFHQ256 using the official implementation. For DPS, we also reuse the official implementation with the available pretrained model, which was trained on FFHQ. For SR3, since no official implementation is available, we used an open-source (non-official) implementation (Jiang, 2022), and we trained a model on FFHQ. When training SR3, we found that a color shift (Deck and Bischoff, 2023) was responsible for important reconstruction errors. To compensate this weakness of the method, we project the super-resolved image on the space of consistent solutions at inference as proposed in (Bahat and Michaeli, 2020). For fair comparison, we retrained both HCFlow and SR3 with the same computational budget as for our low-resolution encoder. For HCFlow and CVDVAE, we set the temperature of the latent variables at $\tau = 0.8$ during sampling.

Evaluating a Diverse SR Method. Due to the ill-posedness of the problem, evaluating a diverse super-resolution model based solely on the distortion to the ground truth is not satisfactory. Indeed, there exist many solutions that are both realistic and consistent with the LR input while being far from the ground truth. Thus, in order to evaluate the super-resolution model, we provide a series of metrics that evaluate different expected characteristics of a diverse super-resolution model, such as the consistency of the solution, the diversity of the samples and the general visual quality. It should be noted that those metrics are not necessarily correlated: a model could generate diverse solutions, that are not consistent or realistic, or, on the opposite, it could provide solutions that are realistic and consistent but with a low diversity. Thus, to evaluate a diverse super-resolution model, it is necessary to consider these three different aspects together: diversity, consistency and visual quality.

Evaluation Metrics. The general quality of the super-resolved images is evaluated using the blind Image quality metric BRISQUE (Mittal et al., 2012). Consistency with the LR input is measured via peak Signal-to-Noise Ratio (PSNR, denoted as LR-PSNR in Tables 1 and 2). Furthermore, to evaluate the diversity of the super-resolution, we compute the Average Pairwise distance between different samples coming from the same LR input (denoted as APD in Tables 1 and 2), both at the pixel level, using the mean square error (MSE) between samples (considering pixel intensity value between 0 and 1), and at a perceptual level using the LPIPS similarity criteria (Zhang et al., 2018). For one LR input, the average pairwise distance is computed as the average distance between all the possible pairs of images in a set of 5 super-resolved samples. The reported APD in Tables 1 and 2 corresponds to the mean value of the single image APD over 500 LR inputs in the test set. We measure the distortion of the super-resolved samples with respect to the ground truth HR image in terms of PSNR, structural similarity (SSIM) (Wang et al., 2004) and LPIPS, as it is common in the super-resolution literature. All numbers reported correspond to the metric mean value on a subset of 1000 images from FFHQ256 test set.

4.2 Results

Quantitative Evaluation. The quantitative results presented in Table 1 indicate that CVDVAE provides a good trade-off between the different evaluated metrics. Indeed, it obtains the second best results in terms of distortion and visual quality, and the second or third best results in terms of diversity. CVDVAE is also one of the fastest methods, along with HCFlow. HCFlow provides the best results for distortion metrics as it explicitly penalizes bad reconstruction in its training loss. Similar to CVDVAE, its application is fast, as it requires only one network evaluation to produce a super-resolved image. However, HCFlow lacks high-level diversity (as measured by the LPIPS average pairwise distance), compared with the concurrent methods. We postulate that this lack of diversity is due to the relative lack of expressiveness of normalizing flows architecture compared to the convolutional architectures used by diffusion and HVAE models. Our method, along with DPS, produces the best

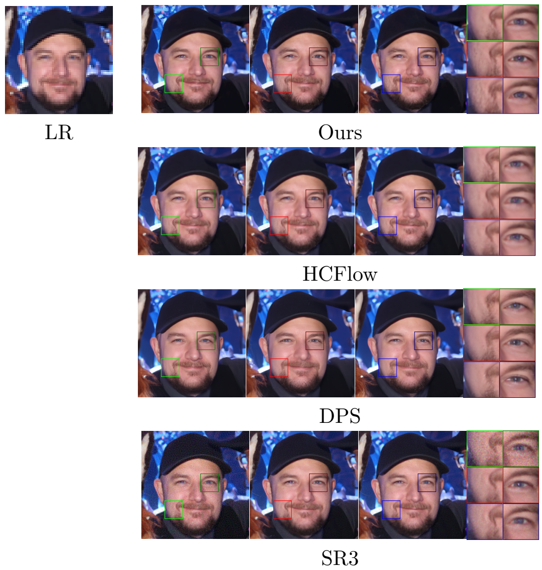


Figure 3: Samples from different diverse SR methods ($\times 4$).

results in terms of visual quality as measured by the BRISQUE metric, illustrating the benefit of using a pretrained unconditional generative model. The computational cost of DPS is nevertheless significantly higher than the ones of CVDVAE and HCFlow, as DPS requires 1000 steps of network evaluations (and backpropagation through the denoiser) to produce one super-resolved sample. Finally, SR3 performances are inferior to the compared methods. We used the same computational budget (48 hours on 4 GPUs) for training the SR3 models than our CVDVAE and HCFlow. This computational budget is significantly lower than the one reported in the SR3 paper (Saharia et al., 2021) (≈ 4 days on 64 TPUv3 chip), and we expect that training the SR3 model for more epochs would improve its performance. Like DPS, SR3 is slower than our method as it requires 2000 network evaluations to produce one super-resolved image; although, unlike DPS, SR3 does not require to backpropagate through the score-network.

Qualitative Evaluation. Visual comparisons of super-resolved samples from the different evaluated methods are provided in Figures 3 and 4. CVDVAE is able to produce diverse textures as illustrated by the facial hair variation in Figure 3 or the hair variation in 4. CVDVAE appears to produce super-resolved samples with higher semantic diversity, in terms of textures (hairs, skin), in line with the higher perceptual diversity measured in the quantitative evaluation.

Temperature Control. As for the unconditionnal HVAE models, CVDVAE offers the possibility to control the conditional generation via the tempera-

Table 1: Comparison of diverse SR methods face super-resolution. Best result is in bold, second best is underlined.

		Distortion			Visual Quality	Consistency	Diversity (APD)		
	model	PSNR↑	SSIM↑	LPIPS ↓	BRISQUE↓	LR-PSNR ↑	MSE ($\times 10^4$) ↑	LPIPS ($\times 10^3$) ↑	time (s)
×4	Bicubic	27.49	0.84	0.29	61.79	36.99	0	0	
	HCFlow	**31.74**	**0.89**	**0.13**	37.21	52.81	**161.8**	62.6	**0.11**
	CVDVAE (ours)	<u>30.24</u>	<u>0.85</u>	<u>0.16</u>	<u>32.30</u>	**75.20**	<u>88.8</u>	123.0	<u>0.14</u>
	SR3	28.87	0.73	0.25	37.17	<u>63.47</u>	20.06	**209.2**	46
	DPS	28.50	0.81	0.20	**32.21**	38.96	10.4	<u>150.0</u>	103
×8	Bicubic	23.50	0.70	0.45	78.42	33.61	0	0	
	HCFlow	**26.72**	**0.76**	**0.24**	36.25	51.13	**575.5**	155.3	<u>0.17</u>
	Ours	25.47	<u>0.71</u>	<u>0.27</u>	<u>32.26</u>	**70.15**	<u>248.2</u>	<u>236.4</u>	**0.13**
	SR3	<u>26.26</u>	0.70	0.29	34.78	<u>68.6</u>	19.95	234.3	62
	DPS	24.38	0.68	0.28	**30.09**	36.97	35.68	**247.4**	103

ture of the latent variable distributions (Vahdat and Kautz, 2020; Child, 2021). The temperature parameter τ controls the variance of the Gaussian latent distributions. In order to assess the behavior of the model on both low and high temperature regime, we evaluate our method on 2 temperatures ($\tau \in \{0.1, 0.8\}$). Quantitative results in Table 2 show that reducing the temperature leads to a solution closer to the ground truth in terms of low-levels distortion metrics (PSNR and the SSIM), while using a higher temperature helps to improve the perceptual similarity (LPIPS) with the ground-truth, as well as the general perceptual quality of the generated HR images and the diversity of the samples. On Figure 5, we display CVDVAE's samples at different temperatures τ. The sampling temperature correlates with the perceptual smoothness of the super-resolved sample, a higher sampling temperature inducing images with sharper details.

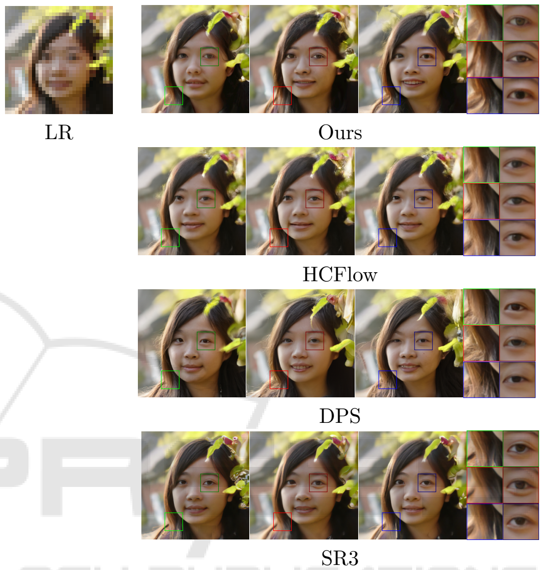


Figure 4: Samples from different diverse SR methods (×8.)

5 RELATED WORKS

Super-Resolution with Pre-Trained Generative Models A large number of methods were designed to solve imaging inverse problems such as image super-resolution by using pretrained deep generative models (DGM) as a prior. This includes methods relying on generative adversarial networks (GAN) (Menon et al., 2020; Marinescu et al., 2020; Pan et al., 2021; Daras et al., 2021; Daras et al., 2022; Poirier-Ginter and Lalonde, 2023), variational autoencoders (Mattei and Frellsen, 2018; González et al., 2022; Prost et al., 2023) and denoising diffusion models (Choi et al., 2021; Chung et al., 2022; Kawar et al., 2022; Song et al., 2023). However, those approaches are computationally expensive as they require an iterative sampling or optimization procedure which require many network evaluation. On the other hand, our approach enables fast inference (one network evaluation), at the cost of reduced flexibility (due to the need of training a task-specific encoder). The idea of training an encoder to map a degraded image in the latent space of a generative network was previously exploited in the context of image inpainting with HVAE (Harvey et al., 2022), and super-resolution with a GAN prior (Chan et al., 2021; Richardson et al., 2021).

Diverse Super-Resolution with Conditional Generative Models. Although it is possible to sample from the posterior $p(\boldsymbol{x}|\boldsymbol{y})$ by using an unconditionnal deep generative models, those methods are restricted to specific dataset for which pretrained models are available. On generic natural images, the state of the art methods rely on conditional generative models, directly trained to model the posterior $p(\boldsymbol{x}|\boldsymbol{y})$ (Lugmayr et al., 2021; Lugmayr et al., 2022). Those methods include conditional normalizing flows (Lugmayr et al., 2020; Liang et al., 2021), conditional GAN (Bahat and Michaeli, 2020), conditional VAE (Gatopoulos et al., 2020; Zhou et al., 2021; Chira et al., 2022) and conditional denoising diffusion models (Li et al., 2022; Saharia et al., 2021).

Table 2: Effect of the sampling temperature τ on CVDVAE super-resolution results.

		Distortion			Visual Quality	Consistency	Diversity (APD)	
	τ	PSNR↑	SSIM↑	LPIPS ↓	BRISQUE↓	LR-PSNR ↑	MSE ($\times 10^4$) ↑	LPIPS ($\times 10^3$) ↑
×4	0.1	**30.75**	**0.86**	**0.15**	36.47	**75.70**	64.6	104.5
	0.8	30.24	0.85	0.16	**32.3**	75.20	**88.8**	**123.0**
×8	0.1	**26.27**	**0.75**	0.30	50.34	**71.63**	140.4	179.0
	0.8	25.47	0.708	**0.28**	32.26	70.15	248.2	236.4

LR $\tau = 0$ $\tau = 0.25$ $\tau = 0.5$ $\tau = 0.75$ $\tau = 1$

Figure 5: Effect of the sampling temperature τ on the super-resolved result. Increasing the temperature yields image with more high-frequency details.

6 CONCLUSIONS

In this work we presented CVDVAE, a method that realizes an efficient sampling from the posterior of a super-resolution problem, by combining a low-resolution image encoder with a pretrained VD-VAE generative model. CVDVAE showed promising results on face super-resolution, on par with state-of-the-art diverse SR methods, providing semantically diverse and high-quality samples. Our results illustrate the ability of conditional hierarchical generative models to perform complex image-to-image tasks.

Our results are in line with many works that illustrates the benefits of using HVAE models for downstream applications (Havtorn et al., 2021; Agarwal et al., 2023; Prost et al., 2023). One drawback of our approach is its limitation to dataset for which pretrained HVAE models are available, such as human faces or low-resolution ImageNet. However, we postulate that HVAE models have not yet reached their limits, and, by adapting design features from current SOTA deep generative models (Rombach et al., 2022; Kang et al., 2023) (architectural improvement, longer training, larger dataset), HVAE models could significantly improve their performance and expreessiveness, and generalize on much diverse datasets.

ACKNOWLEDGEMENTS

This study has been carried out with financial support from the French Research Agency through the PostProdLEAP project (ANR-19-CE23-0027-01). Computer experiments for this work ran on several platforms including HPC resources from GENCI-IDRIS (Grant 2021-AD011011641R1), and the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see https://www.plafrim.fr).

REFERENCES

Agarwal, S., Hope, G., Younis, A., and Sudderth, E. B. (2023). A decoder suffices for query-adaptive variational inference. In *The 39th Conference on Uncertainty in Artificial Intelligence*.

Bahat, Y. and Michaeli, T. (2020). Explorable super resolution. In *the IEEE/CVF CVPR*, pages 2716–2725.

Chan, K. C., Wang, X., Xu, X., Gu, J., and Loy, C. C. (2021). Glean: Generative latent bank for large-factor image super-resolution. In *IEEE/CVF CVPR*, pages 14245–14254.

Child, R. (2021). Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:12011.10650*.

Chira, D., Haralampiev, I., Winther, O., Dittadi, A., and Liévin, V. (2022). Image super-resolution with deep variational autoencoders.

Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S. (2021). Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. (2022). Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh ICLR*.

Daras, G., Dagan, Y., Dimakis, A., and Daskalakis, C. (2022). Score-guided intermediate level optimization: Fast langevin mixing for inverse problems. In *ICML*, pages 4722–4753. PMLR.

Daras, G., Dean, J., Jalal, A., and Dimakis, A. (2021). Intermediate layer optimization for inverse problems using deep generative models. In *ICML*, pages 2421–2432. PMLR.

Deck, K. and Bischoff, T. (2023). Easing color shifts in score-based diffusion models. *arXiv preprint arXiv:2306.15832*.

Gatopoulos, I., Stol, M., and Tomczak, J. M. (2020). Super-resolution variational auto-encoders. *arXiv preprint arXiv:2006.05218*.

González, M., Almansa, A., and Tan, P. (2022). Solving inverse problems by joint posterior maximization with autoencoding prior. *SIAM Journal on Imaging Sciences*, 15(2):822–859.

Harvey, W., Naderiparizi, S., and Wood, F. (2022). Conditional image generation by conditioning variational auto-encoders. In *ICLR*.

Havtorn, J. D., Frellsen, J., Hauberg, S., and Maaløe, L. (2021). Hierarchical vaes know what they don't know. In *ICML*, pages 4117–4128. PMLR.

Hazami, L., Mama, R., and Thurairatnam, R. (2022). Efficient-vdvae: Less is more.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851.

Jiang, L. (2022). Image super-resolution via iterative refinement. https://github.com/Janspiry/Image-Super-Resolution-via-Iterative-Refinement.

Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., and Park, T. (2023). Scaling up gans for text-to-image synthesis. In *IEEE/CVF CVPR*, pages 10124–10134.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *IEEE/CVF CVPR*, pages 4401–4410.

Kawar, B., Elad, M., Ermon, S., and Song, J. (2022). Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lepcha, D. C., Goyal, B., Dogra, A., and Goyal, V. (2022). Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*.

Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., and Chen, Y. (2022). Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59.

Liang, J., Lugmayr, A., Zhang, K., Danelljan, M., Van Gool, L., and Timofte, R. (2021). Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *IEEE/CVF ICCV*, pages 4076–4085.

Lugmayr, A., Danelljan, M., Gool, L. V., and Timofte, R. (2020). Srflow: Learning the super-resolution space with normalizing flow. In *European conference on computer vision*, pages 715–732. Springer.

Lugmayr, A., Danelljan, M., and Timofte, R. (2021). Ntire 2021 learning the super-resolution space challenge. In *IEEE/CVF CVPR*, pages 596–612.

Lugmayr, A., Danelljan, M., Timofte, R., Kim, K.-w., Kim, Y., Lee, J.-y., Li, Z., Pan, J., Shim, D., Song, K.-U., et al. (2022). Ntire 2022 challenge on learning the super-resolution space. In *IEEE/CVF CVPR*, pages 786–797.

Marinescu, R. V., Moyer, D., and Golland, P. (2020). Bayesian image reconstruction using deep generative models. *arXiv preprint arXiv:2012.04567*.

Mattei, P.-A. and Frellsen, J. (2018). Leveraging the exact likelihood of deep latent variable models. *NeurIPS*, 31.

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. (2020). Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *IEEE/CVF CVPR*, pages 2437–2445.

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Trans. on Image Processing*, 21(12):4695–4708.

Pan, X., Zhan, X., Dai, B., Lin, D., Loy, C. C., and Luo, P. (2021). Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 44(11):7474–7489.

Poirier-Ginter, Y. and Lalonde, J.-F. (2023). Robust unsupervised stylegan image restoration. In *IEEE/CVF CVPR*, pages 22292–22301.

Prost, J., Houdard, A., Almansa, A., and Papadakis, N. (2023). Inverse problem regularization with hierarchical variational autoencoders. In *IEEE/CVF ICCV*, pages 22894–22905.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., and Cohen-Or, D. (2021). Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF CVPR*, pages 2287–2296.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *IEEE/CVF CVPR*, pages 10684–10695.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. (2021). Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*.

Sønderby, C. K., Raiko, T., Maalø e, L., Sø nderby, S. r. K., and Winther, O. (2016). How to train deep variational autoencoders and probabilistic ladder networks. In *NeurIPS*, volume 29.

Song, J., Vahdat, A., Mardani, M., and Kautz, J. (2023). Pseudoinverse-Guided Diffusion Models for Inverse Problems. In *(ICLR) ICLR*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *ICLR*.

Vahdat, A. and Kautz, J. (2020). Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE CVPR*, pages 586–595.

Zhou, H., Huang, C., Gao, S., and Zhuang, X. (2021). Vspsr: Explorable super-resolution via variational sparse representation. In *IEEE/CVF CVPR*, pages 373–381.