Inverse problem regularization with hierarchical variational autoencoders

Jean Prost
Univ. Bordeaux, CNRS, Bordeaux INP
IMB, UMR 5251, F-33400 Talence, France
jean.prost@math.u-bordeaux.fr

Antoine Houdard
Ubisoft La Forge
Bordeaux
antoine.houdard@ubisoft.com

Andrés Almansa
Université Paris Cité, CNRS
MAP5, UMR 8145
andres.almansa@parisdescartes.fr

Nicolas Papadakis
Univ. Bordeaux, CNRS, Bordeaux INP
IMB, UMR 5251, F-33400 Talence, France
nicolas.papadakis@math.u-bordeaux.fr

Abstract

In this paper, we propose to regularize ill-posed inverse problems using a deep hierarchical variational autoencoder (HVAE) as an image prior. The proposed method synthesizes the advantages of i) denoiser-based Plug & Play approaches and ii) generative model based approaches to inverse problems. First, we exploit VAE properties to design an efficient algorithm that benefits from convergence guarantees of Plug-and-Play (PnP) methods. Second, our approach is not restricted to specialized datasets and the proposed PnP-HVAE model is able to solve image restoration problems on natural images of any size. Our experiments show that the proposed PnP-HVAE method is competitive with both SOTA denoiser-based PnP approaches, and other SOTA restoration methods based on generative models. The code for this project is available at https://github.com/jprost76/PnP-HVAE.

1. Introduction

In this work, we study linear inverse problems

$$\boldsymbol{y} = A\boldsymbol{x} + \epsilon \tag{1}$$

in which $\boldsymbol{y} \in \mathbb{R}^m$ is the degraded observation, $\boldsymbol{x} \in \mathbb{R}^d$ the original signal we wish to retrieve, $A \in \mathbb{R}^{m \times d}$ is an observation matrix and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is an additive Gaussian noise. Many image restoration tasks can be formulated as (1), including deblurring, super-resolution or inpainting.

With the development of deep learning in computer vision, image restoration have known significant progress. The most straight-forward way to exploit deep learning for solving image inverse problems is to train a neural network to map degraded images to their clean version in a supervised fashion. However, this type of approach requires a large amount of training data, and it lacks flexibility, as one network is needed for each different inverse problem.

An alternate approach is to use deep latent variable generative models such as GANs or VAEs and to compute the Maximum-a-Posterior (MAP) estimator in the latent space:

$$\hat{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} \log p\left(\boldsymbol{y}|G(\boldsymbol{z})\right) + \log p\left(\boldsymbol{z}\right), \tag{2}$$

where $\boldsymbol{z}$ is the latent variable and G is the generative network [4, 34]. In (2) the likelihood $p\left(\boldsymbol{y}|G(\boldsymbol{z})\right)$ is related to the forward model (1), and $p\left(\boldsymbol{z}\right)$ corresponds to the prior distribution over the latent space. After solving (2), the solution of the inverse problem is defined as $\hat{\boldsymbol{x}} = G\left(\hat{\boldsymbol{z}}\right)$. The latent optimization methods (2) provide high-quality solutions that are guaranteed to be in the range of a generative network. However, this implies highly non-convex problems (2) due to the complexity of the generator and the obtained solutions may lack of consistency with the degraded observation [47]. Although the convergence of latent optimization algorithms has been studied in the literature, existing convergence guarantees are either restricted to specific settings, or rely on assumptions that are hard to verify.

In this work, we propose an algorithm that exploits the strong prior of a deep generative model while providing realistic convergence guarantees. We consider a specific type of deep generative model, the hierarchical variational autoencoder (HVAE). HVAE gives state-of-the-art results on image generation benchmarks [53, 6, 15, 31], and provides an encoder that will be key in the design of our proposed method.

As the HVAE model differs significantly from the architecture of concurrent models, it is necessary to design algorithms adapted to their specific structure. The latent space dimension of HVAE is significantly higher than the image dimension. Hence, constraining the solution to lie in the image of the generator is not sufficient enough to regularize inverse problems. Indeed, it has been observed that HVAEs can perfectly reconstruct out-of-domain images [14]. Consequently, we propose to constrain the latent variable of the solution to lie in the high probability area of the HVAE prior distribution. This can be done efficiently by controlling the variance of the prior over the latent variables.

The common practice of optimizing the latent variables of the generative model with backpropagation is impractical due to the high dimensionality of the hierarchical latent space. Instead, we exploit the HVAE encoder to define an alternating algorithm [11] to optimize the joint distribution over the image and its latent variable

To derive convergence guarantees for our algorithm, we show that it can be reformulated as a Plug-and-Play (PnP) method [54], which alternates between an application of the proximal operator of the data-fidelity term, and a reconstruction by the HVAE. Under this perspective, we give sufficient conditions to ensure the convergence of our method, and we provide an explicit characterization of the fixed-point of the iterations. Motivated by the parallel with PnP methods, we name our method PnP-HVAE.

1.1. Contributions and outline

In this work, we introduce PnP-HVAE, a method for regularizing image restoration problems with a hierarchical variational autoencoder. Our approach exploits the expressiveness of a deep HVAE generative model and its capacity to provide a strong prior on specialized datasets, as well as convergence guarantees of Plug-and-Play methods and their ability to deal with natural images of any size. After a review of related works (section 2) and of the background on HVAEs (section 3), our contributions are the following.

- In section 4, we introduce PnP-HVAE, an algorithm to solve inverse problems with a HVAE prior. PnP-HVAE optimizes a joint posterior on image and latent variables without backpropagation through the generative network. It can be viewed as a generalization of JPMAP [11] to hierarchical VAEs, with additional control of the regularization.
- In section 5, we demonstrate the convergence of PnP-HVAE under hypotheses on the autoencoder reconstruction. Numerical experiments illustrate that the technical hypotheses are empirically met on noisy images with our proposed architecture. We also exhibit the better convergence properties of our alternate algorithm with respect to the use of Adam for optimizing the joint posterior objective.
- In section 6, we demonstrate the effectiveness of PnP-HVAE through image restoration experiments and comparisons on (i) faces images using the pre-trained VDVAE model from [6]; and (ii) natural images using the proposed PatchVDVAE architecture trained on natural image patches.

2. Related works

CNN methods for regularizing inverse problems can be classified in two categories: regularization with denoisers (Plug-and-Play) and regularization with generative models.

2.1. Plug-and-Play methods

Plug-and-Play (PnP) and RED methods [54, 43] make use of a (deep) denoising method as a proxy to encode the local information over the prior distribution. The denoiser is plugged in an optimization algorithm such as Half-Quadratic Splitting or ADMM in order to solve the inverse problem. PnP algorithms come with theoretical convergence guarantees by imposing certain conditions on the denoiser network [45, 37, 19]. These approaches provide state-of-the-art results on a wide variety of image modality thanks to the excellent performance of the currently available deep denoiser architectures [55]. However, PnP methods are only implicitly related to a probabilistic model, and they provide limited performance for challenging structured problems such as the inpainting of large occlusions.

2.2. Deep generative models for inverse problems

Generative models represent an explicit image prior that can be used to regularize ill-posed inverse problems [4, 26, 34, 8, 35, 36, 51]. They are latent variable models parametrized by neural networks, optimized to fit a training data distribution [25, 12, 10, 16].

Convergence issues. Regularization with generative models (2) involves highly non-convex optimization over latent variables [4, 34, 35]. The convergence guarantees of existing methods remain to be established [48, 40, 17].

Convergent methods have only been proposed for restricted uses cases. In compressed sensing problems with Gaussian measurement matrices, one can show that the objective function has a few critical points and design an algorithm to find the global optimum [13, 18]. With a prior given by a VAE, and under technical hypothesis on the encoder and the VAE architecture, the Joint Posterior Maximization with Autoencoding Prior (JPMAP) framework of [11] converges toward a minimizer of the joint posterior $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y})$ of the image $\boldsymbol{x}$ and latent $\boldsymbol{z}$ given the observation $\boldsymbol{y}$. JPMAP is nevertheless only designed for VAEs of limited expressiveness, with a simple fixed Gaussian prior distribution over the latent space. This makes it impossible to use this approach for anything other than toy examples.

Genericity issues. When a highly structured dataset with fixed image size is available (*e.g.* face images [20]), deep generative models produce image restorations of impressive quality for severely ill-posed problems, such as super-resolution with huge upscaling factors.

For natural images, deep priors have been improved by normalizing flows [41, 10], GANs [12], score-based [52] and diffusion models [16, 47, 46]. Note that the use of diffusion models in PnP is still based on approximations [50, 7], assumptions [22, 33] or empirical algorithms [30].

While GAN models were considered SOTA, modern hierarchical VAE (HVAE) architectures were shown to display quality on par with GANs [6, 53], while being able to perfectly reconstruct out-of-domain images [14].

Integrating HVAE in a convergent scheme for natural image restoration of any size raises several theoretical and methodological challenges, as the image model of HVAE is the push-forward of a *causal cascade of latent distributions*.

3. Background on variational autoencoders

This work exploits the capacity of HVAEs to model complex image distributions [6, 53]. We review the properties of the loss function used to train a VAE (section 3.1), then we detail the generalization of VAE to hierarchical VAE (section 3.2), and present the temperature scaling approach to monitor the quality of generated images (section 3.3).

3.1. VAE training

Variational autoencoders (VAE) have been introduced in [25] to model complex data distributions. VAEs are trained to fit a parametric probability distribution in the form of a latent variable model:

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z}) p_\theta(\boldsymbol{z}) d\boldsymbol{z}, \tag{3}$$

where $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is a probabilistic decoder, and $p_\theta(\boldsymbol{z})$ corresponds to the prior distribution over the model latent variable $\boldsymbol{z}$. A VAE is also composed of a probabilistic encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, whose role is to approximate the posterior of the latent model $p_\theta(\boldsymbol{z}|\boldsymbol{x})$, which is usually intractable.

The generative model parameters $\theta \in \Theta$ and the approximate posterior parameters $\phi \in \Phi$ of a VAE are jointly trained by maximizing the evidence lower bound (ELBO) [25, 42] on a training data distribution $p_{\text{data}}(\boldsymbol{x})$.

$$\mathcal{L}(\boldsymbol{x}; \theta, \phi) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) || p_\theta(\boldsymbol{z})). \tag{4}$$

The ELBO expectation on $p_{\text{data}}(\boldsymbol{x})$ is upper-bounded by the negative entropy of the data distribution, and, when the upper-bound is reached we have that [56]:

$$\mathrm{KL}(p_{\text{data}}(\boldsymbol{x}) q_\phi(\boldsymbol{z}|\boldsymbol{x}) || p_\theta(\boldsymbol{x}) p_\theta(\boldsymbol{z}|\boldsymbol{x})) = 0. \tag{5}$$

3.2. Hierarchical variational autoencoders

The ability of hierarchical VAEs to model complex distributions is due to their hierarchical structure imposed in the latent space. The latent variable of HVAE is partitioned into L subgroups $\boldsymbol{z} = (\boldsymbol{z}_0, \boldsymbol{z}_1, \cdots, \boldsymbol{z}_{L-1})$, and the prior and the encoder are respectively defined as:

$$p_\theta(\boldsymbol{z}) = \prod_{l=1}^{L-1} p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l}) p_\theta(\boldsymbol{z}_0) \tag{6}$$

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \prod_{l=1}^{L-1} q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{<l}, \boldsymbol{x}) q_\phi(\boldsymbol{z}_0|\boldsymbol{x}), \tag{7}$$

We consider a specific class of HVAEs with Gaussian conditional distributions for the encoder and the decoder

$$\begin{cases} p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l}) &= \mathcal{N}(\boldsymbol{z}_l; \mu_{\theta,l}(\boldsymbol{z}_{<l}) \Sigma_{\theta,l}(\boldsymbol{z}_{<l})) \\ q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{<l}, \boldsymbol{x}) &= \mathcal{N}(\boldsymbol{z}_l; \mu_{\phi,l}(\boldsymbol{z}_{<l}, \boldsymbol{x}), \Sigma_{\phi,l}(\boldsymbol{z}_{<l}, \boldsymbol{x})), \end{cases} \tag{8}$$

where $\mu_{\theta,0}$ and $\Sigma_{\theta,0}$ can either be trainable or non-trainable constants, and the remaining mean vectors ($\mu_{\theta,l}$, and $\mu_{\phi,l}$, for $l > 0$) and covariance matrices ($\Sigma_{\theta,l}$ and $\Sigma_{\phi,l}$, for $l > 0$) are parametrized by neural networks. [1] In this work, we consider models with a Gaussian decoder:

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{x}; \mu_\theta(\boldsymbol{z}), \gamma^2 I\right). \tag{9}$$

3.3. Temperature scaling

As demonstrated in [53, 6], sampling the latent variables $\boldsymbol{z}_l$ from a prior with reduced temperature improves the visual quality of the generated images from the VAE. In practice, this is done by multipling the covariance matrix of the Gaussian distribution $p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l})$ by a factor $\tau_l < 1$. This factor τ_l is called temperature because of its link to statistical physics. Reducing the temperature of the priors amounts to defining the auxilliary model:

$$p_{\theta,\boldsymbol{\tau}}(\boldsymbol{z}_0, \cdots, \boldsymbol{z}_{L-1}, \boldsymbol{x}) = \prod_{l=0}^{L-1} \frac{p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l})^{\frac{1}{\tau_l^2}}}{\tau_l^{\frac{d_l}{2}}} p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<L}), \tag{10}$$

where $\boldsymbol{\tau} := (\tau_0, \cdots, \tau_{L-1})$ gives the temperature for each level of the hierarchy, and d_l is the dimension of the latent variable $\boldsymbol{z}_l$. In the following, we use this temperature-scaled model to balance the regularization of our inverse problem.

4. Regularization with HVAE Prior

In this section we introduce our Plug-and-Play method using a Hierarchical VAE prior (PnP-HVAE) to solve generic image inverse problems. Building on top of the JPMAP framework [11], we propose a joint model over the image and its latent variable that we optimize in an alternate way. By doing so, we take advantage of the HVAE encoder to avoid backpropagation through the generative network. Although our motivation is similar to JPMAP, PnP-HVAE

[1] Note that for the special case $l = 0$, $\boldsymbol{z}_{<l}$ is empty, meaning that $\Sigma_{\theta,0}(\boldsymbol{z}_{<0}) = \Sigma_{\theta,0}$ is actually a constant, $q_\phi(\boldsymbol{z}_0|\boldsymbol{z}_{<0}, \boldsymbol{x}) = q_\phi(\boldsymbol{z}_0|\boldsymbol{x})$ is only conditioned on $\boldsymbol{x}$, etc.

overcomes two of its limitations, that are the lack of control, and the limitation to simple and non-hierarchical VAEs. We show in section 4.1 that the strength of the regularization of the tackled inverse problem can be monitored by tuning the temperature of the prior in the latent space. In section 4.2, we propose an approximation of the joint posterior distribution using the hierarchical VAE encoder. In section 4.3, we present our final algorithm that includes a new greedy scheme to optimize the latent variable of the HVAE.

4.1. Tempered hierarchical joint posterior

The linear image degradation model (1) yields

$$p(\boldsymbol{y}|\boldsymbol{x}) \propto e^{-f(\boldsymbol{x})}, \quad f(\boldsymbol{x}) = \frac{1}{2\sigma^2}||A\boldsymbol{x} - \boldsymbol{y}||^2. \tag{11}$$

Solving the underlying image inverse problem in a bayesian framework requires an a priori distribution $p(\boldsymbol{x})$ over clean images and studying the posterior distribution of $\boldsymbol{x}$ knowing its degraded observation $\boldsymbol{y}$. In this work, the image prior is given by a hierarchical VAE and we exploit the joint posterior model $p(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{y})$. From the HVAE latent variable model with reduced temperature $p_{\theta,\tau}\left(\boldsymbol{z}_0, \cdots, \boldsymbol{z}_{l-1}, \boldsymbol{x}\right)$ as defined in (10), we define the associated tempered joint model as:

$$p\left(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y}\right) := p_{\theta,\tau}\left(\boldsymbol{z}_0, \cdots, \boldsymbol{z}_{L-1}, \boldsymbol{x}\right) p\left(\boldsymbol{y}|\boldsymbol{x}\right). \tag{12}$$

Following the JPMAP idea from [11], we aim at finding the couple $(\boldsymbol{x}, \boldsymbol{z})$ that maximizes the joint posterior $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y})$:

$$\min_{\boldsymbol{x},\boldsymbol{z}} -\log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y}). \tag{13}$$

Although we are only interested in finding the image $\boldsymbol{x}$, the joint Maximum A Posteriori (MAP) criterion (13) makes it possible to derive an optimization scheme that only relies on forward calls of the HVAE. Using Bayes' rule and the definition of the tempered HVAE joint model (12) and (10), the logarithm of the joint posterior rewrites:

$$\begin{aligned} &\log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{y}) \log p(\boldsymbol{y}) \\ &= \log p(\boldsymbol{y}|\boldsymbol{x}) + \sum_{l=0}^{L-1} \log \frac{p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l})^{\frac{1}{\tau_l^2}}}{\tau_l^{\frac{d_l}{2}}} + \log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<L}). \end{aligned} \tag{14}$$

Since $p(\boldsymbol{y})$ is constant, finding the joint MAP estimate (13) amounts to minimizing the following criterion:

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{z}} J_1\left(\boldsymbol{x}, \boldsymbol{z}\right) := &-\sum_{l=1}^{L-1} \frac{1}{\tau_l^2} \log p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l}) \\ &+ f(\boldsymbol{x}) - \log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<L}). \end{aligned} \tag{15}$$

Notice that the temperature of the prior over the latent space τ_l controls the weight of the regularization over the latent variable $\boldsymbol{z}_l$. Optimizing (15) w.r.t. $\boldsymbol{x}$ is tractable, whereas the minimization w.r.t.$\boldsymbol{z}$ requires a backpropagation through the decoder $\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_{<L})$ that is impractical due to the high dimensionality and the hierarchical structure of the HVAE latent space.

4.2. Encoder approximation of the joint posterior

Using the encoder q_ϕ, we can reformulate the joint MAP problem (15) in a form that is more convenient to optimize with respect to $\boldsymbol{z}$. Indeed, assuming that the encoder perfectly matches the true posterior, we have that:

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x}) p_{\text{data}}(\boldsymbol{x})}{p_\theta(\boldsymbol{z})}. \tag{16}$$

This assumption can be met if the variational family $\{q_\phi(.|\boldsymbol{x}); \phi \in \Phi\}$ contains the true posterior $p(\boldsymbol{z}|\boldsymbol{x})$ and is trained to optimality, following equation (5). If this assumption appears unrealistic for vanilla (non-hierarchical) VAE [11], our experiments suggest that HVAE is sufficiently expressive to match the posterior to a reasonably good accuracy.

Thus, by introducing the decoder expression (16) in the full model (12), we have:

$$p\left(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y}\right) = \prod_{l=0}^{L-1} \frac{1}{\tau^{\frac{d_l}{2}}} \frac{q_\phi(\boldsymbol{z}_l|\boldsymbol{z}_{<l}, \boldsymbol{x})}{p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l})^{1-\tau_l^{-2}}} p(\boldsymbol{y}|\boldsymbol{x}) p_{\text{data}}(\boldsymbol{x}).$$

Denoting $\lambda_l = \frac{1}{\tau_l^2} - 1$, we reformulate the joint MAP problem (13) as a joint MAP problem over the encoder model:

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{z}} J_2(\boldsymbol{x}, \boldsymbol{z}) := &-\sum_{l=0}^{L-1} (\log q_\phi(\boldsymbol{z}_l|\boldsymbol{x}, \boldsymbol{z}_{<l}) + \lambda_l \log p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l})) \\ &+ f(\boldsymbol{x}) - \log p_{\text{data}}(\boldsymbol{x}). \end{aligned} \tag{17}$$

4.3. Alternate optimization with PnP-HVAE

We introduce an alternate scheme to minimize (13) that sequentially optimizes with respect to $\boldsymbol{x}$ and to $\boldsymbol{z}$. For a linear degradation model and a Gaussian decoder (1), the criterion $J_1(\boldsymbol{x}, \boldsymbol{z})$ in (15) is convex in $\boldsymbol{x}$ and its global minimum is $\boldsymbol{x} = \left(A^t A + \frac{\sigma^2}{\gamma^2} \mathrm{Id}\right)^{-1} \left(A^t \boldsymbol{y} + \frac{\sigma^2}{\gamma^2} \mu_\theta\left(\boldsymbol{z}\right)\right)$. Next we propose to compute an approximate solution of the problem $\min_{\boldsymbol{z}} J_2(\boldsymbol{x}, \boldsymbol{z})$ with the greedy algorithm 1.

Algorithm 1 Hierarchical encoding with latent regularization to minimize (17) w.r.t. $\boldsymbol{z}$ for a fixed $\boldsymbol{x}$

Require: image $\boldsymbol{x}$; HVAE (ϕ, θ); temperature τ_l; $\lambda_l = \frac{1}{\tau_l^2} - 1$
  for $0 \leq l < L$ **do**
    $S_q \leftarrow \Sigma_{\phi,l}^{-1}\left(\boldsymbol{z}_{<l}, \boldsymbol{x}\right)$; $m_q \leftarrow \mu_{\phi,l}\left(\boldsymbol{z}_{<l}, \boldsymbol{x}\right)$ ▷ Encoder
    $S_p \leftarrow \Sigma_{\theta,l}^{-1}\left(\boldsymbol{z}_{<l}\right)$; $m_p \leftarrow \mu_{\theta,l}\left(\boldsymbol{z}_{<l}\right)$ ▷ Prior
    $\boldsymbol{z}_l \leftarrow \left(S_q + \lambda_l S_p\right)^{-1}\left(S_q m_q + \lambda_l S_p m_p\right)$
  end for
  return $E_\tau\left(\boldsymbol{x}\right) = \left(\hat{\boldsymbol{z}}_0, \hat{\boldsymbol{z}}_1, \cdots, \hat{\boldsymbol{z}}_{L-1}\right)$

In algorithm 1, the latent variables $\boldsymbol{z}_l$ are determined in a hierarchical fashion starting from the coarsest to the finest one. As defined in relations (8), the conditionals $q_\phi(\boldsymbol{z}_l|\boldsymbol{x}, \boldsymbol{z}_{<l})$ and $p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l})$ are Gaussian. Therefore, the minimization at each step can be viewed as an interpolation between the mean of the encoder $q_\phi(\boldsymbol{z}_l|\boldsymbol{x}, \boldsymbol{z}_{<l})$ and the prior $p_\theta(\boldsymbol{z}_l|\boldsymbol{z}_{<l})$, given some weights conditioned by the covariance matrices and the temperature τ_l. The solution of each minimization problem in $\boldsymbol{z}_l$ is given by:

$$\hat{z}_l = \left(\Sigma_{\phi,l}^{-1}\left(\hat{\boldsymbol{z}}_{<l}, \boldsymbol{x}\right) + \lambda_l \Sigma_{\theta,l}^{-1}\left(\hat{\boldsymbol{z}}_{<l}\right)\right)^{-1} \left(\Sigma_{\phi,l}^{-1}\left(\hat{\boldsymbol{z}}_{<l}, \boldsymbol{x}\right)\mu_{\phi,l}\left(\hat{\boldsymbol{z}}_{<l}, \boldsymbol{x}\right) + \lambda_l \Sigma_{\theta,l}^{-1}\left(\hat{\boldsymbol{z}}_{<l}\right)\mu_{\theta,l}\left(\hat{\boldsymbol{z}}_{<l}\right)\right) \tag{18}$$

where $\lambda_l = 1/\tau_l^2 - 1$. In the following, we denote as $\hat{\boldsymbol{z}} := E_{\boldsymbol{\tau}}(\boldsymbol{x})$ the output of the hierarchical encoding of Algorithm 1. In appendix 8.1 we show that this algorithm finds the global optimum of $J_2(\boldsymbol{x}, \cdot)$ under mild assumptions.

The final PnP-HVAE procedure to solve an inverse problem with the HVAE prior is presented in Algorithm 2.

Algorithm 2 PnP-HVAE - Restoration by solving (15)

> $k \leftarrow 0$; $res \leftarrow +\infty$; initialize $\boldsymbol{x}^{(0)}$
> **while** $res > tol$ **do**
> $\quad$ % $\min_{\boldsymbol{z}} J_2(\boldsymbol{x}^{(k)}, \boldsymbol{z})$ $\quad\triangleright$ Optimize (17) w.r.t. $\boldsymbol{z}$ using Alg. 1
> $\quad \boldsymbol{z}^{(k+1)} = E_{\boldsymbol{\tau}}(\boldsymbol{x}^{(k)})$
> $\quad$ % $\min_{\boldsymbol{x}} J_1(\boldsymbol{x}, \boldsymbol{z}^{(k+1)})$ $\quad\triangleright$ Optimize (15) w.r.t. $\boldsymbol{x}$
> $\quad \boldsymbol{x}^{(k+1)} = \left(A^tA + \frac{\sigma^2}{\gamma^2}\,\mathrm{Id}\right)^{-1}\left(A^t\boldsymbol{y} + \frac{\sigma^2}{\gamma^2}\mu_\theta\left(\boldsymbol{z}^{(k+1)}\right)\right)$
> $\quad res \leftarrow ||\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}||$; $k \leftarrow k+1$
> **end while**
> **return** $\boldsymbol{x}^{(k)}$

5. Convergence analysis

We now analyse the convergence of Algorithm 2. Following the work of [3], the alternate optimization scheme converges if $q_\phi(\boldsymbol{z}|x) = p_\theta(\boldsymbol{z}|\boldsymbol{x})$ and the greedy optimization scheme in Algorithm 1 actually solves $\min_{\boldsymbol{z}} J_2(\boldsymbol{x}, \boldsymbol{z})$. In practice, it is difficult to verify if these hypotheses hold. We propose to theoretically study algorithm 2, and next verify empirically that the assumptions are met.

In section 5.1, we reformulate Algorithm 2 as a Plug-and-Play algorithm, where the HVAE reconstruction takes the role of the denoiser. Then we study in section 5.2 the fixed-point convergence of the algorithm. Finally, section 5.3 contains numerical experiments with the patch HVAE architecture proposed in section 6.2. We empirically show that the patch architecture satisfies the aforementioned technical assumptions and then illustrate the numerical convergence and the stability of our alternate algorithm.

5.1. Plug-and-Play HVAE

In this section we make the assumption that the HVAE decoder is Gaussian with a constant variance on its diagonal (9). We rely on the proximal operator of a convex function f that is defined as $\mathrm{prox}_f(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} f(\boldsymbol{u}) + \frac{1}{2}||\boldsymbol{x} - \boldsymbol{u}||^2$.

Proposition 1 *Assume the decoder is defined as in (9). Denote HVAE$(\boldsymbol{x}, \boldsymbol{\tau}) := \mu_\theta(E_{\boldsymbol{\tau}}(\boldsymbol{x}))$. Then the alternate scheme described in Algorithm 2 writes*

$$\boldsymbol{x}_{k+1} = \mathrm{prox}_{\gamma^2 f}\left(\mathit{HVAE}\left(\boldsymbol{x}_k, \boldsymbol{\tau}\right)\right). \tag{19}$$

From relation (19), algorithm 2 is a Plug-and-Play Half-Quadratic Splitting method [45] where the role of the denoiser is played by the reconstruction HVAE$(\boldsymbol{x}_k, \boldsymbol{\tau})$. We now derive from relation (19) sufficient conditions to establish the convergence of the iterations.

5.2. Fixed-point convergence

Let us denote T the operator corresponding to one iteration of (19): $\mathrm{T}(\boldsymbol{x}) = \mathrm{prox}_{\gamma^2} f\left(\mathrm{HVAE}\left(\boldsymbol{x}, \boldsymbol{\tau}\right)\right)$. The Lipschitz constant of T can then be expressed as a function of f and the HVAE reconstruction operator HVAE$(\boldsymbol{x}_k, \boldsymbol{\tau})$.

Proposition 2 (Proof in supplementary) *Assume that the decoder has a constant variance $\Sigma_\theta^{-1}(\boldsymbol{z}) = \frac{1}{\gamma^2}\,\mathrm{Id}$ for all $\boldsymbol{z}$; and the autoencoder with latent regularization is $L_{\boldsymbol{\tau}}$-Lipschitz,* i.e. $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^{\mathbf{n}}$: $||\mathit{HVAE}(\mathbf{u}, \boldsymbol{\tau}) - \mathit{HVAE}(\mathbf{v}, \boldsymbol{\tau})|| \leq L_{\boldsymbol{\tau}}||\mathbf{u} - \mathbf{v}||$. *Then, denoting as $\lambda_{\min}$ the smallest eigenvalue of A^tA, we have*

$$||T(\mathbf{u}) - T(\mathbf{v})|| \leq \frac{\sigma^2}{\gamma^2\lambda_{\min} + \sigma^2} L_\tau ||\mathbf{u} - \mathbf{v}||. \tag{20}$$

Corollary 1 *If HVAE$(\boldsymbol{x}_k, \boldsymbol{\tau})$ is $L_{\boldsymbol{\tau}} < 1$-Lipschitz, then iterations (19) converge.*

Proof. If $L_{\boldsymbol{\tau}} < 1$, then HVAE$(\boldsymbol{x}_k, \boldsymbol{\tau})$ is a contraction. Hence T is also a contraction form proposition 2 and consequently, Banach theorem ensures the convergence of the iteration $\boldsymbol{x}_{k+1} = T(\boldsymbol{x}_k)$ to a fixed point of T. □

Proposition 3 (Proof in supplementary) *$\boldsymbol{x}^\star$ is a fixed point of T if and only if:*

$$\nabla f(\boldsymbol{x}^\star) = \frac{1}{\gamma^2}\left(\mathit{HVAE}\left(\boldsymbol{x}^\star, \boldsymbol{\tau}\right) - \boldsymbol{x}^\star\right) \tag{21}$$

Proposition 3 characterizes the solution of the latent-regularization scheme, in the case where the HVAE reconstruction is a contraction. Under mild assumptions, the fixed point condition can be stated as a critical point condition

$$\nabla f(\boldsymbol{x}^*) + \nabla g(\boldsymbol{x}^*) = 0,$$

of the objective function $f(\boldsymbol{x}) + g(\boldsymbol{x}) = -\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p_{\theta,\boldsymbol{\tau}}(\boldsymbol{x})$, where the tempered prior is the marginal $p_{\theta,\boldsymbol{\tau}}(\boldsymbol{x}) := \int p_{\theta,\boldsymbol{\tau}}(\boldsymbol{x}, \boldsymbol{z})d\boldsymbol{z}$ of the joint tempered prior defined in (10). This result, detailed in the supplementary, follows from an interpretation of HVAE$(\boldsymbol{x}, \boldsymbol{\tau})$ as an MMSE denoiser. As a consequence Tweedie's formula provides the link between the right-hand side of equation (21) and ∇g.

5.3. Numerical convergence with PatchVDVAE

We illustrate the numerical convergence of Algorithm 2. We first analyse the Lipschitz constant of the HVAE reconstruction with the PatchVDVAE architecture proposed in section 6.2. Then we study the empirical convergence of the algorithm and show that it outperforms the baseline optimisation of the joint MAP (15) with the Adam optimizer.

Lipschitz constant of the HVAE reconstruction. Corollary 1 establishes the fixed point convergence of our proposed optimization algorithm under the hypothesis that the reconstruction with latent regularization is a contraction, *i.e.* $L_{\boldsymbol{\tau}} < 1$. We now show thanks to an empirical estimation of the Lipschitz constant $L_{\boldsymbol{\tau}}$ that our PatchVDVAE network empirically satisfies such a property when applied to noisy images. We present in figure 1 the histograms of the ratios $r = ||\mathrm{HVAE}(\mathbf{u}, \boldsymbol{\tau}) - \mathrm{HVAE}(\mathbf{v}, \boldsymbol{\tau})||/||\mathbf{u} - \mathbf{v}||$, where $\mathbf{u}$ and $\mathbf{v}$ are natural images extracted from the BSD dataset and corrupted with white Gaussian noise. These ratios give a lower bound for the true Lipschitz constant $L_{\boldsymbol{\tau}}$. Although it is possible to set different temperature τ_l at each level, we fixed a constant temperature amongst all levels to limit the number of hyperparameters. We realized tests for 3 temperatures $\tau \in \{0.6, 0.8, 0.99\}$, and 3 noise levels $\sigma \in \{0, 25, 50\}$. On clean images ($\sigma = 0$), the distribution of ratios in close to 1. This suggests that the HVAE is well trained and accurately models clean images. In some rare case, a ratio $r \geq 1$ is observed for clean images. This indicates that the reconstruction is not a contraction everywhere, in particular on the manifold of clean images.

On noisy images $\sigma > 0$, the reconstruction behaves as a contraction, as the ratio $r < 1$ is always observed. Moreover, reducing the temperature of the latent regularization τ increases the strength of the contraction. This suggests that with the trained PatchVDVAE architecture, the hypothesis $L_{\boldsymbol{\tau}} < 1$ in Corollary 1 holds for noisy images.

Convergence of Algorithm 2 We now illustrate the effectiveness of PnP-HVAE through comparisons with the optimization of the objective $J_1(\boldsymbol{x}_k, \boldsymbol{z}_k)$ in (15) using the Adam algorithm [23] for two learning rates $lr \in \{0.01, 0.001\}$. The left plot in figure 2 shows that Adam is able to estimate a better minimum of J_1. However, our alternate algorithm requires a smaller number of iterations to converge.

On the other hand, as illustrated by the right plot in figure 2, the use of Adam involves numerical instabilities. Oscillations of the ratio $L_k := \frac{||\mathrm{T}(\boldsymbol{x}_{k+1}) - \mathrm{T}(\boldsymbol{x}_k)||}{||\boldsymbol{x}_{k+1} - \boldsymbol{x}_k||}$ are even increased with larger learning rates, whereas our method provides a stable sequence of iterates.

More important, we finally exhibit the better quality of the restorations obtained with our alternate algorithm on inpaiting, deblurring and super-resolution of face images. In these experiments, we used the hierarchical VDVAE model [6] trained on the FFHQ dataset [20]. Figure 3 (see 2nd and 4th columns) and table 1 (PSNR, SSIM and LPIPS scores) illustrate that the quality of the images restored with our alternate optimization algorithm is higher than the ones obtained with Adam. This suggests that for image restoration purposes, our optimization method is able to find a more relevant local minimum of J_1 than Adam.

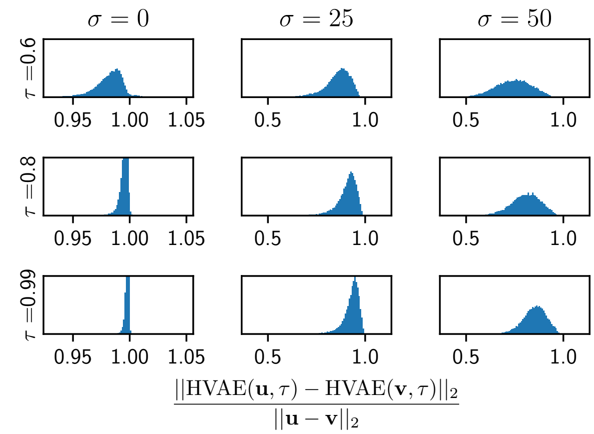


Figure 1: Numerical estimation of the Lipschitz constant of PatchVDVAE reconstruction with different temperatures τ. We present the histogram of ratio values $\frac{||\mathrm{HVAE}(\mathbf{u},\tau) - \mathrm{HVAE}(\mathbf{v},\tau)||}{||\mathbf{u}-\mathbf{v}||}$, where $\mathbf{u}$ and $\mathbf{v}$ are natural images corrupted with white Gaussian noise of different standard deviations σ. For noisy images ($\sigma > 0$), the observed Lipschitz constant is always less than 1.

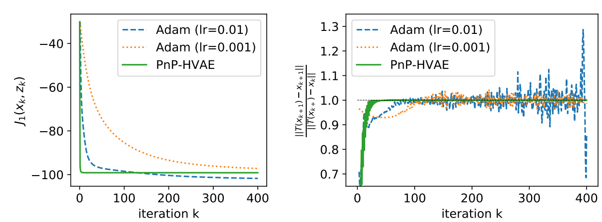


Figure 2: Comparison of the convergence of PnP-HVAE algorithm 2 with respect to the baseline Adam optimizer, on a deblurring problem. Left (Convergence of the function value): PnP-HVAE converges faster to a minimum of the joint posterior $J_1(\boldsymbol{x}_k, \boldsymbol{z}_k)$ in (15). Right (Convergence of iterates $\boldsymbol{x}_k$): PnP-HVAE is more stable than Adam.

6. Image restoration results

We present in section 6.1 an application of PnP-HVAE on face images, using a pretrained state-of-the-art hierarchical VAE. Next, we study the application of our framework to natural images. To that end, we introduce in section 6.2 a patch hierachical VAE architecture, that is able to model natural images of different resolutions. In section 6.3, we provide deblurring, super-resolution and inpainting experiments to demonstrate the relevance of the proposed method.

Figure 3: Visual comparaison of image restoration methods based on deep generative models. We studied 3 tasks on face images: inpainting (top), deblurring (middle), super-resolution (bottom). Contrary to the optimization of the objective (15) with Adam, our alternate algorithm generates realistic results, on par with ILO [8], while remaining consistant with the observation.

		PSNR↑	SSIM↑	LPIPS↓	time (s)
SR $\times 4$ $\sigma = 3$	Adam	28.56	0.75	0.38	26
	ILO	28.80	0.78	0.17	34
	PnP-HVAE	**29.32**	**0.82**	0.28	**15**
	DPS	27.53	0.76	**0.12**	153
Deblurring (motion) $\sigma = 8$	Adam	26.69	0.75	0.27	12
	ILO	29.01	0.80	0.20	34
	PnP-HVAE	**30.40**	**0.84**	0.16	**10**
	DPS	28.70	0.80	**0.11**	142
Deblurring (Gaussian) $\sigma = 8$	Adam	30.17	0.83	0.21	12
	ILO	29.12	0.79	0.17	34
	PnP-HVAE	**30.81**	**0.86**	0.24	**10**
	DPS	29.14	0.81	**0.10**	142

Table 1: Quantitative evalutation on face restoration. Best results in **bold**, second best underlined.

6.1. Face Image restoration (FFHQ)

We first demonstrate the effectiveness of PnP-HVAE on highly structured data, by performing face image restoration. Latent variable generative models can accurately model structured images such as face images [20, 53, 6, 24], and then be used to produce high quality restoration of such data. In our experiments, we use the VDVAE model of [6], pre-trained on the FFHQ dataset [20], as our hierarchical VAE prior. VDVAE has $L = 66$ latent variable groups in its hierarchy and generates images at resolution 256×256.

We compare PnP-HVAE with two restoration methods based on different class of generative models, namely the intermediate layer optimization algorithm (ILO) [8] and the diffusion posterior sampling method (DPS) [7]. ILO is a GAN inversion method which optimizes the image latent code along with the intermediate layer representation of a StyleGAN2 generative network [21] to generate an image consistent with a degraded observation. DPS use denoising diffusion probabilistic model [52, 16] as a prior, and produce a sample from the posterior by conditioning each iteration of the sampling process on $\boldsymbol{y}$. We use the official implementation of ILO, along with a StyleGAN2 model that was trained for 550k iterations on images of resolution 256×256 from FFHQ [44]. For DPS, we use the official implementation as well. As VDVAE and StyleGAN models are not trained on the same train-test split of FFHQ, we chose to evaluate the methods on a subset of 100 images from the CelebA dataset [29]. For super-resolution, the degradation model corresponds to the application of a Gaussian low-pass filter followed by a $\times 4$ sub-sampling, and the addition of a Gaussian white noise with $\sigma = 3$. For the deblurring, we considered motion blur and Gaussian kernels, both with a noise level $\sigma = 8$.

We provide quantitative comparisons in table 1, along with a visual comparison of the results in figure 3. PnP-HVAE has the best PSNR and SSIM results for all the considered restoration tasks, while ILO provides better results for the perceptual distance. By jointly optimizing the image and its latent variable, PnP-HVAE provides results that are both realistic and consistent with the degraded observation. On the other hand, ILO only optimizes on an extended latent space. This method generates sharp and realistic images with better LPIPS scores, but the results lack of consistency with respect to the observation, which explains the overall lower PSNR performance. DPS produces highly realistic samples, and gives the best perceptual distance. However DPS is limited by its long inference time, as it requires one network function evaluation and one backpropagation operation through the network at each of the 1000 sampling steps required to generate one image.

6.2. PatchVDVAE: a HVAE for natural images

Available generative models in the literature operate on images of fixed resolutions and are either restrained to datasets of limited diversity, or even to registered face images [24, 6, 53, 20], or requiring additional class information [5, 9, 52, 31]. Fitting an unconditional model on natural images appears to be a more difficult task, as their resolution can change, and their content is highly diverse. The complexity of the problem can be reduced by learning a prior model on patches of reduced dimension. For image restoration problems, the patch model can be reused on images of higher dimensions [57, 39, 2]. When the model is a full CNN, the prior on the set of the patches can be computed efficiently by applying the network on the full image [39].

We thus introduce patchVDVAE, a fully convolutional hierarchical VAE. Contrary to existing HVAE models whose resolution is constrained by the constant tensor at the input of the top-down block, patchVDVAE can generate images of different resolutions by controlling the dimension of the input latent. This amounts to defining a prior on patches whose dimension corresponds to the receptive field of the VAE. A similar model is used for image denoising in [38].

For PatchVDVAE architecture, we use the same bottom-up and top-down blocks as VDVAE [6], and replace the constant trainable input in the first top-down block by a latent variable, to make the model fully convolutional (details on the architecture are given in the supplementary). The training dataset is composed of 128×128 patches extracted from a combination of DIV2K [1] and Flickr2K [28] datasets. We perform data augmentation by extracting patches at 3 resolutions: HR-images and $\times 2$ and $\times 4$ downscaled images. The model is trained for 7.10^5 iterations with a batch size of 64. Following the recommendation of [15], we use Adamax optimizer with an exponential moving average and gradient smoothing of the variance. We set the decoder model to be a Gaussian with diagonal covariance, as in [31]. PatchVDVAE is fully convolutional and can generate images of dimension that are multiples of 64 as illustrated by figure 4.

Figure 4: Left: 64×64 patches samples from our patchVDVAE model trained on patches from natural images. Right: PatchVDVAE is fully convolutional and it can generate images of higher resolution (here: 128×128).

6.3. Natural images restoration

We evaluate PnP-HVAE on natural image restoration. For each task, we report the average value of the PSNR, the SSIM, and the LPIPS metrics on 20 images from the test set of the BSD dataset [32].

Image deblurring. In the experiments, we consider 2 Gaussian kernels and 2 motion blur kernels from [27], with 3 different noise levels $\sigma \in \{2.55, 7.65, 12.75\}$. As a baseline we consider EPLL [57], which learns a prior on image patches with a Gaussian mixture model. We also compare PnP-HVAE with PnP-MMO and GS-PnP, 2 competing convergent Plug-and-Play methods based on CNN denoisers. PnP-MMO [37] restricts the denoiser to be contraction in order to guarantee the convergence of the PnP forward-backard algorithm. GS-PnP [19] considers a gradient step denoiser and reaches state-of-the-art performances of non converging methods [55]. We set the temperature τ in our method as 0.95, 0.8 and 0.6 for noise levels 2.55, 7.65 and 12.75 respectively, and we let it run for a maximum of 50 iterations. For the three compared methods we use the official implementations and pre-trained models provided by the respective authors. Details on the choice of hyperparameters for the concurrent methods are provided in the supplementary material. Figure 5 illustrates that our method provides correct deblurring results. According to table 2, the performance of PnP-HVAE is between those of EPLL and GS-PnP and it outperforms PnP-MMO for large noise levels.

σ	Method	PSNR↑	SSIM↑	LPIPS↓
2.55	PnP-HVAE	27.75	0.79	0.31
	GS-PNP [19]	**29.59**	**0.84**	**0.22**
	EPLL [57]	26.49	0.71	0.36
	PnP-MMO [37]	<u>29.50</u>	<u>0.83</u>	<u>0.20</u>
7.65	PnP-HVAE	<u>26.36</u>	<u>0.72</u>	<u>0.40</u>
	GS-PNP [19]	**27.33**	**0.77**	**0.31**
	EPLL [57]	24.04	0.66	0.45
	PnP-MMO [37]	25.34	0.69	0.34
12.75	PnP-HVAE	<u>25.12</u>	**0.73**	<u>0.47</u>
	GS-PNP [19]	**26.32**	**0.73**	**0.37**
	EPLL [57]	23.28	0.61	0.51
	PnP-MMO [37]	22.42	0.53	0.54

Blur and motion kernels

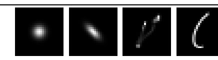
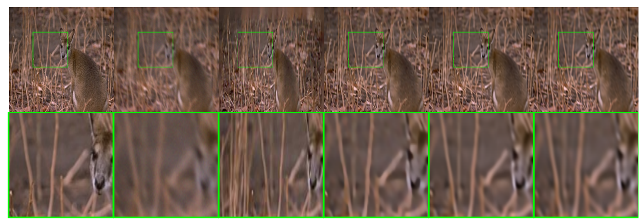
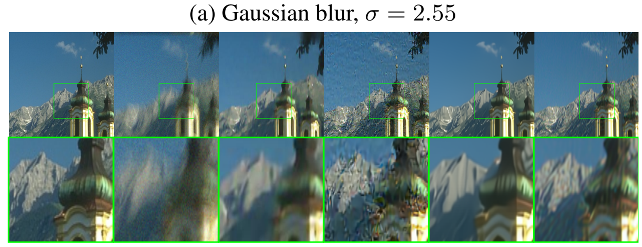


Table 2: Comparison of PnP-HVAE and other restoration methods on deblurring. Results are averaged on 4 kernels.



GT blurry EPLL PnP-MMO GS-PnP PnP-HVAE

(a) Gaussian blur, $\sigma = 2.55$



GT blurry EPLL PnP-MMO GS-PnP PnP-HVAE

(b) Motion blur, $\sigma = 7.65$

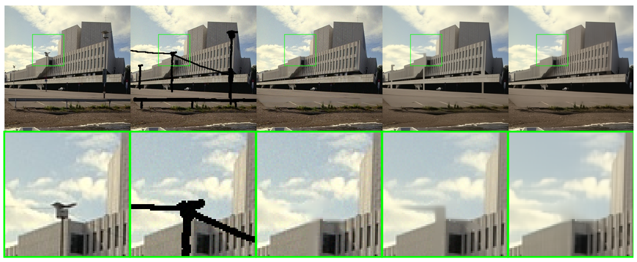
Figure 5: Natural image deblurring

Image inpainting. Next we consider the task of noisy image inpainting. We compose a test-set of 10 images from the validation set of BSD [32] and we create masks by occluding diverse objects of small size in the images. A Gaussian white noise with $\sigma = 3$ is added to the images. As a comparaison, we still consider GS-PnP and EPLL. For PnP-HVAE, the temperature is set to $\tau = 0.6$, and the algorithm is run for a maximum of 200 iterations, unless the residual $||\boldsymbol{x}_{k+1} - \boldsymbol{x}_k||$ is on a plateau. We provide on Table 3 the distortion metrics with the ground truth, as well as a visual comparison on figure 6. With its hierarchical structure,

	PSNR↑	SSIM↑	LPIPS↓
PnP-HVAE	**29.54**	**0.93**	**0.06**
GS-PNP	28.52	**0.93**	0.09
EPLL	<u>29.16</u>	**0.93**	**0.06**

Table 3: Quantitative evaluation for inpainting on BSD.

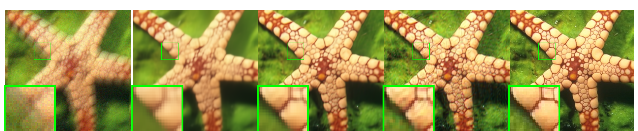
PnP-HVAE outperforms the compared methods.



GT Masked EPLL GS-PnP PnP-HVAE

Figure 6: Natural image inpainting

Effect of the temperature. PnP-HVAE gives control on the temperature of the prior over the latent space. In figure 7, we illustrate that reducing the temperature increases the strength of the regularization prior. In this example the tuning $\tau = 0.7$ produces the best performance.



blurry $\tau = 0.5$ 23.61dB $\tau = 0.7$ 27.32dB $\tau = 0.9$ 26.95dB GT

Figure 7: Effect of the temperature in PnP-VAE on a deblurring problem, with $\sigma = 7.65$.

Effect of the number of latent groups We study the effect of the number of latent groups L on the hierarchical model on the restoration performance. It has been observed that HVAEs outperform non-hierarchical VAEs in terms of likelihood score [49], and that increasing the number of latent groups in the hierarchy improves the modelling performance of HVAE for a fixed number of parameters [6]. Therefore we can expect that the gain in modelling performance due to a higher L will translate into a gain in restoration performance using our method. We train different patchVDVAE models, with different number of latent groups L. In order to keep the number of trainable parameters constant, we replace stochastic top-down blocks with deterministic blocks in our network with the higher L value ($L = 36$). We evaluate the different models on image deblurring, using the same experimental settings as the one described in subsection 6.3. The results in table 4 show that increasing the number of stochastic groups (L) has a positive effect on the validation ELBO, up to $L = 18$, and that a better ELBO correlates with a better restoration performance.

Table 4: Effect of the number L of latent groups on the restoration performance, measured in PSNR (dB), for image deblurring. We observed similar trends for the LPIPS and SSIM metrics.

	$L = 6$	$L = 12$	$L = 18$	$L = 36$
$\sigma = 2.55$	27.25	**27.87**	<u>27.82</u>	27.71
$\sigma = 7.65$	26.10	26.41	**26.74**	<u>26.51</u>
$\sigma = 12.75$	24.78	25.16	**25.57**	<u>25.27</u>
ELBO↑(val)	-1.24	-1.14	$\mathbf{-1.10}$	$\mathbf{-1.10}$

7. Conclusion

We proposed PnP-HVAE, a method using hierarchical variational autoencoders as a prior to solve image inverse problems. Motivated by an alternate optimization scheme, PnP-HVAE exploits the encoder of the HVAE to avoid backpropagating through the generative network. We derived sufficient conditions on the HVAE model to guarantee the convergence of the algorithm. We have verified empirically that PnP-HVAE satisfies those conditions. By jointly optimizing over the image and the latent space, PnP-HVAE produces realistic results that are more consistent with the observation than GAN inversion on a specialized dataset. PnP-HVAE can also restore natural images of any size using our PatchVDVAE model trained on natural images patches.

Acknowledgements

This study has been carried out with financial support from the French Research Agency through the PostProdLEAP project (ANR-19-CE23-0027-01). Experiments presented in this paper were carried out using 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

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 8

[2] Fabian Altekrüger, Alexander Denker, Paul Hagemann, Johannes Hertrich, Peter Maass, and Gabriele Steidl. Patchnr: learning from very few images by patch normalizing flow regularization. *Inverse Problems*, 39(6):064006, 2023. 7

[3] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of operations research*, 35(2):438–457, 2010. 5

[4] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546. PMLR, 2017. 1, 2

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 7

[6] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 6, 7, 8, 9

[7] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2022. 3, 7

[8] Giannis Daras, Joseph Dean, Ajil Jalal, and Alex Dimakis. Intermediate layer optimization for inverse problems using deep generative models. In *International Conference on Machine Learning*, pages 2421–2432. PMLR, 2021. 2, 7

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 7

[10] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2016. 2

[11] Mario González, Andrés Almansa, and Pauline Tan. Solving inverse problems by joint posterior maximization with autoencoding prior. *SIAM Journal on Imaging Sciences*, 15(2):822–859, 2022. 2, 3, 4

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[13] Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[14] Jakob D Havtorn, Jes Frellsen, Søren Hauberg, and Lars Maaløe. Hierarchical vaes know what they don't know. In *International Conference on Machine Learning*, pages 4117–4128. PMLR, 2021. 2, 3

[15] Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficient-vdvae: Less is more. *arXiv preprint arXiv:2203.13751*, 2022. 1, 8

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 7

[17] Matthew Holden, Marcelo Pereyra, and Konstantinos C Zygalakis. Bayesian imaging with data-driven priors encoded by neural networks. *SIAM Journal on Imaging Sciences*, 15(2):892–924, 2022. 2

[18] Wen Huang, Paul Hand, Reinhard Heckel, and Vladislav Voroninski. A provably convergent scheme for compressive sensing under random generative priors. *Journal of Fourier Analysis and Applications*, 27:1–34, 2021. 2

[19] Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations*, 2022. 2, 8

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 6, 7

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 7

[22] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising Diffusion Restoration Models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, volume 2020-Decem, jan 2022. 3

[23] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 6

[24] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 7

[25] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2, 3

[26] Fabian Latorre, Volkan Cevher, et al. Fast and provable admm for learning with generative priors. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[27] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1964–1971. IEEE, 2009. 8

[28] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 8

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 7

[30] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *(CVPR) Conference on Computer Vision and Pattern Recognition*, pages 11451–11461. IEEE, jun 2022. 3

[31] Eric Luhman and Troy Luhman. Optimizing hierarchical image vaes for sample quality. *arXiv preprint arXiv:2210.10205*, 2022. 1, 7, 8

[32] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001. 8

[33] Xiangming Meng and Yoshiyuki Kabashima. Diffusion Model Based Posterior Sampling for Noisy Linear Inverse Problems. nov 2022. 3

[34] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 1, 2

[35] Thomas Oberlin and Mathieu Verm. Regularization via deep generative models: an analysis point of view. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 404–408, 2021. 2

[36] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7474–7489, 2021. 2

[37] Jean-Christophe Pesquet, Audrey Repetti, Matthieu Terris, and Yves Wiaux. Learning maximally monotone operators for image recovery. *SIAM Journal on Imaging Sciences*, 14(3):1206–1237, 2021. 2, 8

[38] Mangal Prakash, Mauricio Delbracio, Peyman Milanfar, and Florian Jug. Interpretable unsupervised diversity denoising and artefact removal. In *International Conference on Learning Representations*, 2021. 8

[39] Jean Prost, Antoine Houdard, Andrés Almansa, and Nicolas Papadakis. Learning local regularization for variational image restoration. In *Scale Space and Variational Methods in Computer Vision: 8th International Conference, SSVM 2021, Virtual Event, May 16–20, 2021, Proceedings*, pages 358–370. Springer, 2021. 7

[40] Ankit Raj, Yuqi Li, and Yoram Bresler. Gan-based projector for faster recovery with convergence guarantees in linear inverse problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5602–5611, 2019. 2

[41] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 2

[42] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014. 3

[43] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM J. on Im. Sc.*, 10(4):1804–1844, 2017. 2

[44] Kim Seonghyeon (rosinality). stylegan2-pytorch. https://github.com/rosinality/stylegan2-pytorch, 2020. 7

[45] Ernest Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, pages 5546–5557. PMLR, 2019. 2, 5

[46] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-Image Diffusion Models. In *(SIGGRAPH) Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, number 1, pages 1–10, New York, NY, USA, aug 2022. ACM. 3

[47] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, apr 2021. 1, 3

[48] Viraj Shah and Chinmay Hegde. Solving linear inverse problems using gan priors: An algorithm with provable guarantees. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4609–4613. IEEE, 2018. 2

[49] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016. 9

[50] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-Guided Diffusion Models for Inverse Problems. In *(ICLR) International Conference on Learning Representations*, 2023. 3

[51] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2021. 2

[52] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020. 2, 7

[53] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020. 1, 3, 7

[54] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013. 2

[55] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021. 2, 8

[56] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from deep generative models. In *International Conference on Machine Learning*, pages 4091–4099. PMLR, 2017. 3

[57] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *2011 international conference on computer vision*, pages 479–486. IEEE, 2011. 7, 8