Image Super-Resolution With Deep Variational Autoencoders

Darius Chira[*,1] **Ilian Haralampiev**[*,1] **Ole Winther**[1,2,3]
Andrea Dittadi[†,1,4] **Valentin Liévin**[†,1]

[1]Technical University of Denmark
[2]University of Copenhagen
[3]Rigshospitalet, Copenhagen University Hospital
[4]Max Planck Institute for Intelligent Systems, Tübingen

Abstract

Image super-resolution (SR) techniques are used to generate a high-resolution image from a low-resolution image. Until now, deep generative models such as autoregressive models and Generative Adversarial Networks (GANs) have proven to be effective at modelling high-resolution images. Models based on Variational Autoencoders (VAEs) have often been criticized for their feeble generative performance, but with new advancements such as VDVAE (very deep VAE), there is now strong evidence that deep VAEs have the potential to outperform current state-of-the-art models for high-resolution image generation. In this paper, we introduce VDVAE-SR, a new model that aims to exploit the most recent deep VAE methodologies to improve upon image super-resolution using transfer learning on pretrained VDVAEs. Through qualitative and quantitative evaluations, we show that the proposed model is competitive with other state-of-the-art methods.

1. Introduction

Single Image Super-Resolution (SISR) consists in producing a high-resolution image from its low-resolution counterpart. Image super-resolution has long been considered one of the most arduous challenges in image processing. This is yet another computer vision task that was transformed by the deep learning revolution and has potential applications including but not limited to medical imaging, security, computer graphics, and surveillance.

Deep generative models have been shown to excel at image generation. This is particularly true for autoregressive models (Chen et al. 2018; Oord et al. 2016; Oord, Kalchbrenner, and Kavukcuoglu 2016; Parmar et al. 2018; Uria et al. 2016) and Generative Adversarial Networks (GANs) (Brock, Donahue, and Simonyan 2018; Goodfellow et al. 2014; Karras, Laine, and Aila 2019; Zhu et al. 2017), whereas Variational Autoencoders (VAEs) (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014) have long been thought to be unable to produce high-quality samples. However, recent improvements in VAE design, such as using a hierarchy of latent variables and increasing depth (Child 2020; Kingma et al. 2016; Maaløe et al. 2019; Vahdat and Kautz 2020) have demonstrated that deep VAEs can compete with both GANs and autoregressive models for high-resolution image generation. The current state-of-the-art VAE is the Very Deep Variational Autoencoder (VDVAE) (Child 2020) which successfully scales to 78 stochastic layers, whereas previous work only experimented with up to 40 layers (Vahdat and Kautz 2020).

[*]Equal contribution, alphabetical order.
[†]Equal advising, alphabetical order. Correspondence to: {adit,valv}@dtu.dk.

Since the VAE is an unconditional generative model, in order to perform image super-resolution it has to be turned into a *conditional* generative model which generates data depending on additional conditioning information. This can be achieved by using the framework of Conditional Variational Autoencoders (CVAE) (Sohn, Lee, and Yan 2015), where the prior is conditioned on an additional variable and parameterized by a neural network. In this work, we introduce **VDVAE-SR**, a VDVAE conditioned on low-resolution images by adding a new component that we call LR-encoder as it resembles the encoder of the original VDVAE. This component is connected to the decoder, passing information on each layer in the top-down path both to the prior and the approximate posterior. During training, the latent distributions of the low- and high-resolution images are matched using the KL divergence term in the ELBO. The learned information is used in generative mode, where only the low-resolution image is included in the model.

A drawback of deep models such as the VDVAE is that they require a large amount of computing and training time. One way to compensate in that regard is to apply transfer learning and utilize a pretrained model in order to speed up the process. However, this is not always straightforward in practice as presenting a pretrained model with new data could lead to exploding gradients. This is particularly relevant for deep variational autoencoders as they are prone to unstable training and can be sensitive to hyperparameters changes. In this paper, we show that using transfer learning for super-resolution on VDVAEs is possible, by making only certain parts of the network trainable and using gate parameters to stabilize the process.

We fine-tune a VDVAE model that was pretrained on FFHQ 256x256 (Karras, Laine, and Aila 2019) using DIV2K (Agustsson and Timofte 2017), a common dataset in the image super-resolution literature (Dai et al. 2019; Lim et al. 2017; Niu et al. 2020; Wang et al. 2018). We evaluate the fine-tuned model on a number of common datasets in the literature of single image super-resolution: Set5 (Bevilacqua et al. 2012), Set14 (Zeyde, Elad, and Protter 2010), Urban100 (Huang, Singh, and Ahuja 2015), BSD100 (Martin et al. 2001), and Manga109 (Matsui et al. 2016). Following previous work (Ledig et al. 2017; Niu et al. 2020; Wang et al. 2018), we test our approach both quantitatively, in terms of PSNR and SSIM metrics, and qualitatively, by visually inspecting the generated images, and compare our results against two state-of-the-art super-resolution methods: EDSR (Lim et al. 2017) and ESRGAN (Wang et al. 2018). We investigate the role of the sampling temperature, which controls the variance of samples at each stochastic layer in VDVAEs, and show results generated with low and high temperatures. By sampling with a lower temperature, the quantitative scores of our model are better than ESRGAN, but slightly worse than EDSR. Qualitatively, sampling with a higher temperature yields sharper images than those generated by the EDSR model. While ESRGAN tends to generate sharper images, we observe that it is prone to produce more unnatural artifacts. We believe that our method achieves a good balance between image sharpness and avoiding unwanted visual artifacts.

We summarize our contributions as follows:

- We propose VDVAE-SR, an adaptation of very deep variational autoencoders (VDVAEs) for the task of single image super-resolution. VDVAE-SR introduces an additional component that we call the LR-encoder, which takes the low-resolution image as input, while its output is used to condition the prior.
- We show how to utilize transfer learning and achieve stable training in order to take advantage of a VDVAE model already pretrained on 32 V100 GPUs for 2.5 weeks.
- We present competitive qualitative and quantitative results compared to state-of-the-art methods on popular test datasets for 4x upscaling. We show how the temperature parameter in VDVAE-SR can be used at test-time for fine-grained control of the trade-off between the sharpness of the generated images and the presence of unnatural artifacts.

2. Related work

One of the first successes in image super-resolution is the SR-CNN (Dong et al. 2015), which is based on a three-layer CNN structure and uses a bicubic interpolated low-resolution image as input to the network. Later, with the proposal of residual neural networks (ResNets) (He et al. 2016), which provide fast training and better performance for deep architectures, numerous works have adapted ResNets-based models to the task of super-resolution, such as SR-ResNet (Ledig et al. 2017) and SR-DenseNet (Tong et al. 2017). One of the frequently used CNN-based super-resolution models in comparative studies is EDSR (Lim et al. 2017), where the authors use ResNets without batch normalization in the residual block, achieving impressive results and getting first place on the NTIRE2017 Super-Resolution Challenge (Timofte et al. 2017).

In terms of GAN-based image super-resolution models, several methods have gained a lot of popularity starting with SRGAN (Ledig et al. 2017) where the authors argue that most popular metrics (PSNR, SSIM) do not necessarily reflect perceptually better SR results and that is why they use an extensive mean opinion score (MOS) for evaluating perceptual quality. With that in mind, SRGAN introduces a perceptual loss different from previous work, based on adversarial as well as content loss. Another method, ESRGAN (Wang et al. 2018), builds upon SRGAN by improving the network architecture removing all batch normalization layers and introducing a new Residual in Residual Dense Block (RRDB). In addition, an enhanced discriminator is used based on Relativistic GAN (Jolicoeur-Martineau 2018) and the features before the activation loss are used to improve perceptual loss.

A recent work that uses VAEs for image super-resolution is the srVAE (Gatopoulos, Stol, and Tomczak 2020), which consists of a VAE with three latent variables, one of them being a downscaled version of the original image. This work shows impressive generative performance in terms of FID score when tested on ImageNet-32 and CIFAR-10, but no quantitative results of their super-resolution model are reported. Another recent work that uses a VAE-based model for image super-resolution is VarSR (Hyun and Heo 2020). This work focuses on very low-resolution images (8x8) and shows better results compared to some popular super-resolution methods.

Deep VAEs (Child 2020; Kingma et al. 2016; Maaløe et al. 2019; Vahdat and Kautz 2020) adapt their architecture from Ladder VAEs (LVAE) (Sønderby et al. 2016), which introduce a novel top-down inference model and achieve stable training with multiple stochastic layers. A method that improved upon the LVAE is the Bidirectional-Inference VAE (BIVA) (Maaløe et al. 2019) adding a deterministic top-down path in the generative model and applying a bidirectional inference network. These modifications solved the variable collapse issue of the LVAE which may occur when the architecture consists of a very deep hierarchy of stochastic latent variables. Recently, NVAE (Vahdat and Kautz 2020) reported further improvements by using normalizing flows in order to allow for more expressive distributions and thus outperform the state-of-the-art among non-autoregressive and VAE models. Finally, the VDVAE model (Child 2020) demonstrated that the number of stochastic layers matters greatly for performance, achieving better results than previous VAE-based models and some autoregressive ones.

Denoising Diffusion Probabilistic Models (DDPM) (Ho, Jain, and Abbeel 2020) are the latest addition to the family of probabilistic generative models. DDPMs define a diffusion process that progressively turns the input image into noise, and learn to synthesize images by inverting that process. DDPMs excel at high-resolution image generation (Dhariwal and Nichol 2021; Nichol and Dhariwal 2021) and have been successfully applied to single image super-resolution (Ho et al. 2022; Li et al. 2022).

3. Preliminaries

In this section, we define variational autoencoders (VAEs) and conditional VAEs (CVAEs), for which we derive the evidence lower bound. We then introduce the VDVAE using the VAE framework.

3.1 Variational autoencoders

The Variational Autoencoder (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014) is a generative model built on probabilistic principles. It consists of a joint model $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x} \,|\, \mathbf{z})p_\theta(\mathbf{z})$ parameterized by θ and an approximate posterior $q_\phi(\mathbf{z} \,|\, \mathbf{x})$ parameterized by ϕ. All models are implemented using neural networks. During generation, the latent variable $\mathbf{z}$ is sampled from the prior and the observation variable $\mathbf{x}$ is sampled from the observation model following $\mathbf{z} \sim p_\theta(\mathbf{z}), \mathbf{x} \sim p_\theta(\mathbf{x} \,|\, \mathbf{z})$.

VAE models are optimized with stochastic gradient ascent to maximize the marginal likelihood:

$$p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x} \,|\, \mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z} \ . \tag{1}$$

In practice, $p_\theta(\mathbf{x})$ is intractable due to the integration over $\mathbf{z}$, which makes the posterior $p_\theta(\mathbf{z} \,|\, \mathbf{x})$ also intractable. Variational Inference (VI) solves the intractability of $p_\theta(\mathbf{z} \,|\, \mathbf{x})$ using an approximate posterior $q_\phi(\mathbf{z} \,|\, \mathbf{x})$. The resulting objective function, the evidence lower bound (ELBO), is further derived using Jensen's inequality and expressed as:

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z} \,|\, \mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \,|\, \mathbf{x})} \right] \leq \log p_\theta(\mathbf{x}) \ . \tag{2}$$

3.2 Conditional variational autoencoders

In order to generate specific data as in the case of image super-resolution, where we need to generate a high-resolution image from its low-resolution counterpart, the Conditional Variational Autoencoders (CVAE) can be used. Similar to the VAE, the CVAE is also built on probabilistic principles. CVAE is optimized to maximize the marginal probability similar to Eq. (2) but this time conditioned on a random variable which could be for example a low-resolution image $\mathbf{y}$:

$$p_\theta(\mathbf{x} \,|\, \mathbf{y}) = \int_{\mathbf{z}} p_\theta(\mathbf{x} \,|\, \mathbf{z}, \mathbf{y}) p(\mathbf{z} \,|\, \mathbf{y}) d\mathbf{z} \ . \tag{3}$$

The posterior of the latent variables is:

$$p_\theta(\mathbf{z} \,|\, \mathbf{x}, \mathbf{y}) = \frac{p_\theta(\mathbf{x} \,|\, \mathbf{z}, \mathbf{y}) p_\theta(\mathbf{z} \,|\, \mathbf{y})}{p_\theta(\mathbf{x} \,|\, \mathbf{y})} \tag{4}$$

where again $p_\theta(\mathbf{x} \,|\, \mathbf{y})$ is intractable, so the posterior has to be approximated using a variational distribution $q_\phi(\mathbf{z} \,|\, \mathbf{x}, \mathbf{y}) \approx p_\theta(\mathbf{z} \,|\, \mathbf{x}, \mathbf{y})$. The conditional ELBO for the CVAE can be derived again using Jensen's inequality, resulting in:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}; \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z} \,|\, \mathbf{x}, \mathbf{y})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z} \,|\, \mathbf{y})}{q_\phi(\mathbf{z} \,|\, \mathbf{x}, \mathbf{y})} \right] \leq \log p_\theta(\mathbf{x} \,|\, \mathbf{y}) \ . \tag{5}$$

3.3 Very Deep Variational Autoencoder (VDVAE)

The VDVAE (Child 2020) consists of a hierarchy of layers of latent variables conditionally dependent on each other. This results in more flexible prior and posterior distributions than simple diagonal Gaussians, which could be too limiting. An iterative interaction between "bottom-up" and "top-down" layers is achieved through parameter sharing between the inference and generative models in each layer. The prior for a model with K stochastic layers factorizes as follows:

$$p_\theta(\mathbf{z}) = p_\theta(\mathbf{z}_0)p_\theta(\mathbf{z}_1 \,|\, \mathbf{z}_0)...p_\theta(\mathbf{z}_K \,|\, \mathbf{z}_{<K}) \tag{6}$$

where $p_\theta(\mathbf{z}_0)$ is a diagonal Gaussian distribution $\mathcal{N}(\mathbf{z}_0 \,|\, \mathbf{0}, \mathbf{I})$ and the latent variable group $\mathbf{z}_0$ is at the top layer, further removed from the data, and typically corresponds to the lowest resolution and smallest number of latent variables. The variable group $\mathbf{z}_K$ is at the bottom of the network and usually has a larger number of latent variables and a high resolution. Similarly, the approximate posterior can be written as follows:

$$q_\phi(\mathbf{z} \,|\, \mathbf{x}) = q_\phi(\mathbf{z}_0 \,|\, \mathbf{x})q_\phi(\mathbf{z}_1 \,|\, \mathbf{z}_0, \mathbf{x})...q_\phi(\mathbf{z}_K \,|\, \mathbf{z}_{<K}, \mathbf{x}) \ . \tag{7}$$

The VDVAE architecture consists of blocks of two types: the residual blocks (bottom-up path) and the top-down blocks (see Fig. 1). The top-down blocks are also residual and handle two tasks: processing the information flowing through the decoder and handling the stochasticity. Each top-down block of index $j > 0$ handles the distributions $q_\phi(\mathbf{z}_j \,|\, \mathbf{z}_{j-1}, \mathbf{x})$ and $p_\theta(\mathbf{z}_j \,|\, \mathbf{z}_{j-1})$. Top-down blocks are composed sequentially. Therefore, we can define $\mathbf{h}_j$ as the input to the top-down block of index j, where $\mathbf{h}_j$ is a function of the samples $\mathbf{z}_{<j}$. This allows us to express the VDVAE model in terms of its prior and variational posterior as:

$$p_\theta(\mathbf{z}) = p_\theta(\mathbf{z}_0) \prod_{j=1}^{K} p_\theta(\mathbf{z}_j \,|\, \mathbf{h}_j) \ , \qquad q_\phi(\mathbf{z} \,|\, \mathbf{x}) = q_\phi(\mathbf{z}_0 \,|\, \mathbf{x}) \prod_{j=1}^{K} q_\phi(\mathbf{z}_j \,|\, \mathbf{h}_j, \mathbf{x}) \ . \tag{8}$$

4. VDVAE-SR

In this section, we introduce the VDVAE-SR model. We provide an overview of the model architecture, after which we detail the conditional prior network and its integration with the VDVAE.

LR-encoder. The dependency on the lower-resolution image $\mathbf{y}$ is implemented using the encoder of a lower-resolution VDVAE of depth $K' < K$, which we call LR-encoder. The LR-encoder maps the lower-resolution image to the latent space, providing one activation $\mathbf{g}_j$ for each layer $j \in [0, K']$. Each activation $\mathbf{g}_j$ is defined as the output of the bottom-up residual block of index j.

Conditional prior. The top-down path, or decoder, of the VDVAE is modified to depend on $\mathbf{y}$ using the LR-encoder activations $\mathbf{g}_0, \ldots, \mathbf{g}_{K'}$. This results in a conditional prior $p_\theta(\mathbf{z} \,|\, \mathbf{y})$ that maps the low-resolution image $\mathbf{y}$ to a distribution over the latent variables $\mathbf{z}$ (see Fig. 2). The architecture of the VDVAE-SR is identical to the one of the VDVAE, except for two alterations:

1. The input to each top-down block (see Fig. 3) is defined as:

$$\tilde{\mathbf{h}}_j = \begin{cases} \mathbf{g}_j & \text{if } j = 0 \\ \mathbf{h}_j + \alpha_j \mathbf{g}_j & \text{if } j \in [1, K'] \\ \mathbf{h}_j & \text{otherwise} \end{cases} \tag{9}$$

where $\alpha_1, \ldots, \alpha_{K'}$ are scalar gate parameters initialized to zero (Bachlechner et al. 2020).

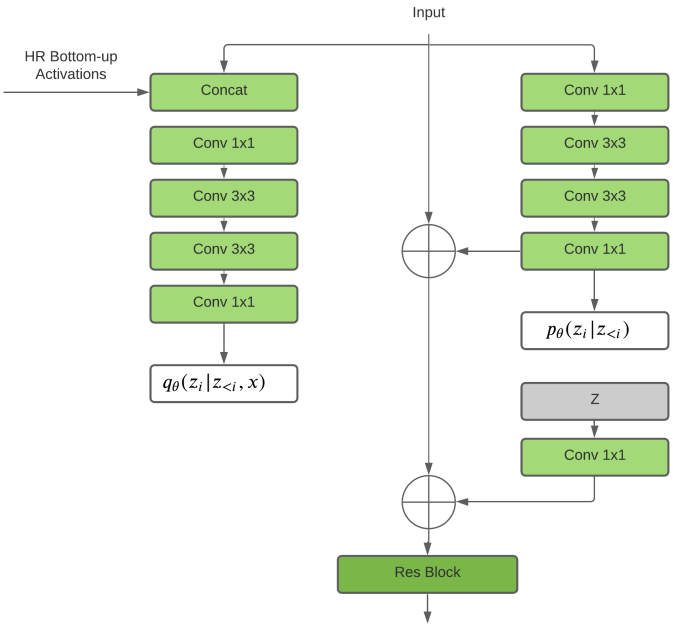

Figure 1: Top-down block of the VDVAE (Child 2020).

2. The top layer is conditioned on $\mathbf{y}$ such that

$$p_\theta(\mathbf{z}_0 \,|\, \tilde{\mathbf{h}}) = \mathcal{N}\left(\mathbf{z}_0 \,|\, \mu_\theta(\tilde{\mathbf{h}}_0),\, \sigma_\theta(\tilde{\mathbf{h}}_0)\right) \ , \tag{10}$$

where μ_θ and σ_θ are linear layers mapping the output of the top-most LR-encoder layer to the parameter-space of $p_\theta(\mathbf{z}_0 \,|\, \mathbf{y})$.

Generative model and inference network. Because of the sharing of the top-down model between the generative model and the inference network (Sønderby et al. 2016), the conditional inference network naturally arises from the alteration of the prior, without further modification. Using the activations $\tilde{\mathbf{h}}_0, \ldots, \tilde{\mathbf{h}}_K$ and the definition of the VDVAE given in Eq. (8), we define the VDVAE-SR as:

$$p_\theta(\mathbf{z} \,|\, \mathbf{y}) = \prod_{j=0}^{K} p_\theta(\mathbf{z}_j \,|\, \tilde{\mathbf{h}}_j) \ , \qquad q_\phi(\mathbf{z} \,|\, \mathbf{y}, \mathbf{x}) = q_\phi(\mathbf{z}_0 \,|\, \mathbf{x}) \prod_{j=1}^{K} q_\phi(\mathbf{z}_j \,|\, \tilde{\mathbf{h}}_j, \mathbf{x}) \ . \tag{11}$$

5. Experiments

5.1 Datasets

Training dataset. We train our models on the DIV2K dataset, introduced by (Agustsson and Timofte 2017). The DIV2K dataset consists of 800 RGB high-definition high-resolution images for training, 100 images for validation, and 100 for testing. The dataset contains a variety of diverse images, including different types of shot such as portrait, scenery, and object shots.

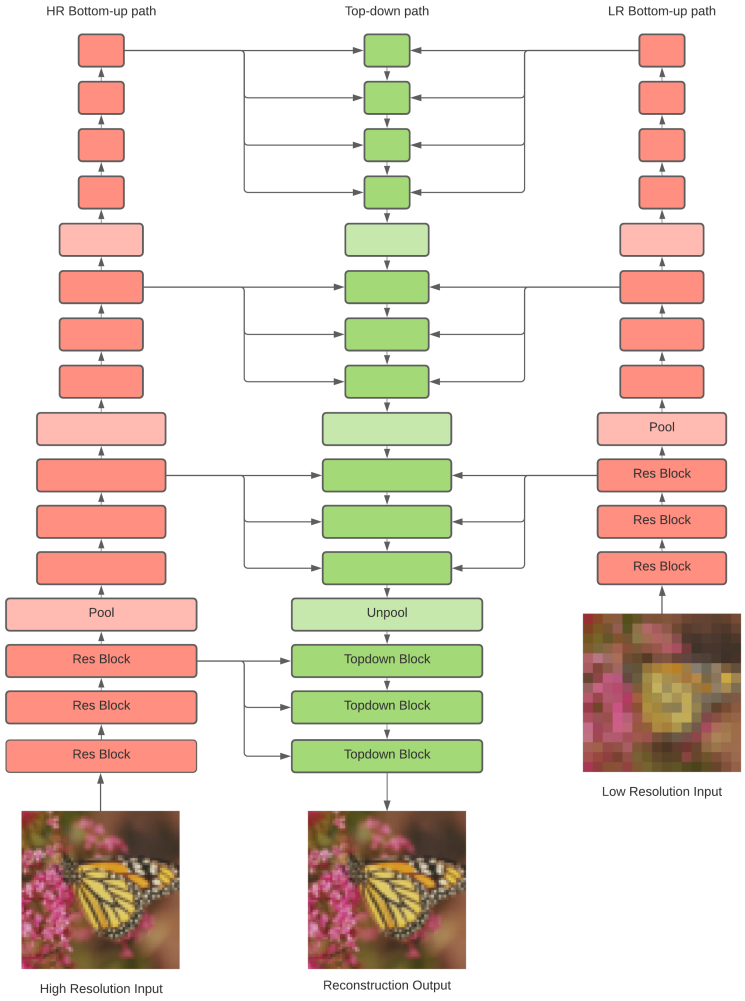

Figure 2: Network architecture of the proposed VDVAE-SR model.

Test datasets. We test our method on popular benchmarking datasets commonly used in single-image super resolution: Set5 (Bevilacqua et al. 2012), Set14 (Zeyde, Elad, and Protter 2010), Urban100 (Huang, Singh, and Ahuja 2015), BSD100 (Martin et al. 2001), and Manga109 (Matsui et al. 2016). Having multiple test datasets gives a better understanding of the strengths and shortcomings of our model, since these datasets contain different types of images: BSD100, Set5, and Set14 mostly consist of natural images with a broad range of styles, while the focus on Urban100 is mainly on buildings and urban scenes, and Manga109 consists of drawings of Japanese manga.

5.2 Implementation details

Since training a VDVAE model on FFHQ 256x256 on 32 NVIDIA V100 GPUs requires about 2.5 weeks, we choose to rely on pretrained VDVAEs and adapt them to the super-resolution task. We use a pretrained VDVAE with a depth of 62 stochastic layers. Our method, VDVAE-SR, includes the original VDVAE encoder and decoder, which we initialize with the weights from the pretrained

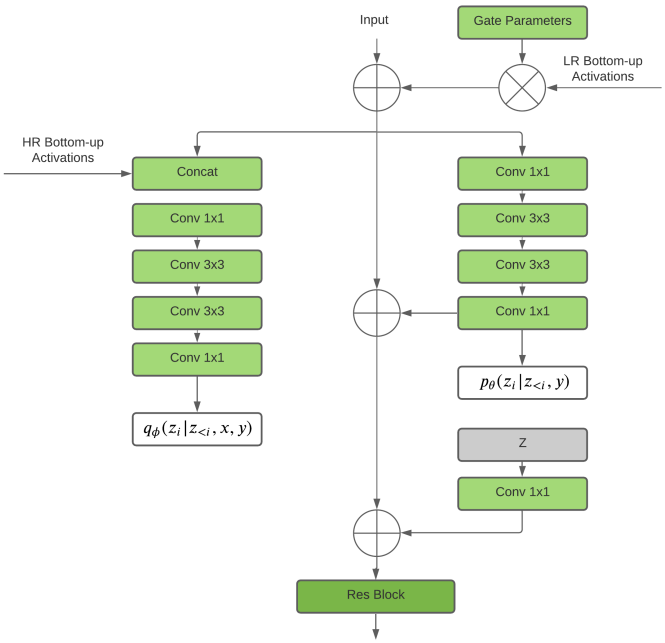

Figure 3: Top-down block of the VDVAE-SR.

model. We then freeze the encoder, allow fine-tuning of the decoder, and train the LR-encoder from scratch. We optimize the model end-to-end for 100,000 steps using the Adam optimizer (Kingma and Ba 2014) with a learning rate of $5 \cdot 10^{-4}$ and a batch size of 1 on one NVIDIA V100 GPU.

When using transfer learning, we observed that the model suffered from exploding gradients if the new information from the LR-encoder was introduced in an uncontrolled manner. Introducing gate parameters similarly to the approach in Bachlechner et al. (2020) significantly improved training stability.

5.3 Evaluation

In terms of evaluation metrics, we use the traditional PSNR and SSIM quality metrics, both widely used for image restoration tasks. While PSNR (Peak Signal to Noise Ratio) is calculated based on the mean squared error of the pixel-to-pixel difference, the SSIM (Structural Similarity Method) is considered to have a closer correlation with human perception by calculating distortion levels based on comparisons of structure, luminance, and contrast. We quantitatively evaluate different super-resolution methods by applying them to low-resolution images and computing the PSNR and SSIM metrics using the super-resolution output and the reference high-resolution image. All images are converted from RGB to YCbCr and the metrics are computed on the Y channel (luma component) of the images. The reason for this is that it has been observed (Pisharoty, Jadhav, and Dandawate 2013)that the results of evaluating on the luminosity channel in the YCbCr color space, rather than on the standard RGB representation, are closer to the actual perceived structural noise of the image. We thus adopt the same approach, following prior work. Finally, note that the YCbCr space is used during the testing phase exclusively, while the training and validation are still performed in the RGB color space.

Table 1: Evaluation metrics using PSNR and SSIM on the Y channel. The number next to VDVAE-SR (our method) denotes the temperature used for sampling. The best scores are represented in **bold**, while the second best results are underlined.

Dataset	EDSR		ESRGAN		VDVAE-SR 0.1		VDVAE-SR 0.8	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Set5	**31.97**	**0.902**	30.39	0.864	31.48	0.886	30.51	0.869
Set14	**28.33**	**0.800**	26.20	0.720	27.99	0.776	27.62	0.761
BSDS100	**28.46**	**0.781**	25.87	0.690	28.05	0.752	27.69	0.738
Manga109	**30.85**	**0.918**	28.77	0.870	29.92	0.904	29.55	0.896
Urban100	**26.02**	**0.798**	24.36	0.748	25.36	0.759	25.15	0.750

5.4 Results

Quantitative results. We compare our method to two other super-resolution methods, namely EDSR and ESRGAN, based on their official implementation. The quantitative results on PSNR and SSIM are shown in Table 1, where EDSR performs best on both metrics, with our method ($t = 0.1$) closely following on second place.

As first discussed in (Ledig et al. 2017), the PSNR and SSIM scores tend to favor smoother images, this being attributed to the nature of how these metrics are calculated, which differs from human visual perception. This is confirmed by the scores obtained by our method using different temperatures, as decreasing the variance produces smoother images and higher scores.

Qualitative results. Figs. 4 to 6 show a visual comparison of two images from BSD100 dataset between the original HR image, Bicubic, EDSR, ESRGAN, and our method with both 0.1 and 0.8 temperatures. It can be observed that the points made in the quantitative section still stand, as EDSR, having the best PSNR and SSIM scores, has a smoother and blurrier look, and our model with 0.1 temperature looks closer to it. As for the model with 0.8 temperature, it introduces more details compared to EDSR. It is still blurrier than the outputs of ESRGAN but has fewer artifacts and it is able to reproduce some details without introducing any generative noise.

In Figs. 4 and 5 it can be observed that ESRGAN produces some artifacts on the bull's head and the person's hand, while our model retains the structure of the objects. In Fig. 6 we can again see how the eye of the bird has a different shape and a more averaged-out look in the case of the EDSR, and even more drastic shape change in the case of the ESRGAN, while our models keep the rounder shape, while not averaging out the outer colors as much.

Temperature. The temperature parameter t, taking values between 0 and 1, is used in VDVAE when sampling from prior in generative mode, often yielding higher-quality samples when decreased (Kingma and Dhariwal 2018; Vahdat and Kautz 2020). Reducing the temperature results in reducing the variance of the Gaussian distributions in the prior, thereby achieving more regularity in the generated samples. Fig. 7 shows examples of super-resolution outputs obtained by sampling with different temperatures. We can observe how samples taken with a lower temperature look smoother, whereas those taken with a higher temperature have more details but also more artifacts. We corroborate this quantitatively in Fig. 8, which shows that the PSNR and SSIM scores (for Set5 and Set14) both decrease as the sampling temperature is increased. This agrees with our qualitative observations, as PSNR and SSIM are usually higher for smoother images that contain less noise.

Figure 4: Comparison of super-resolution models for an image (number 376043) of the BSD100 dataset.

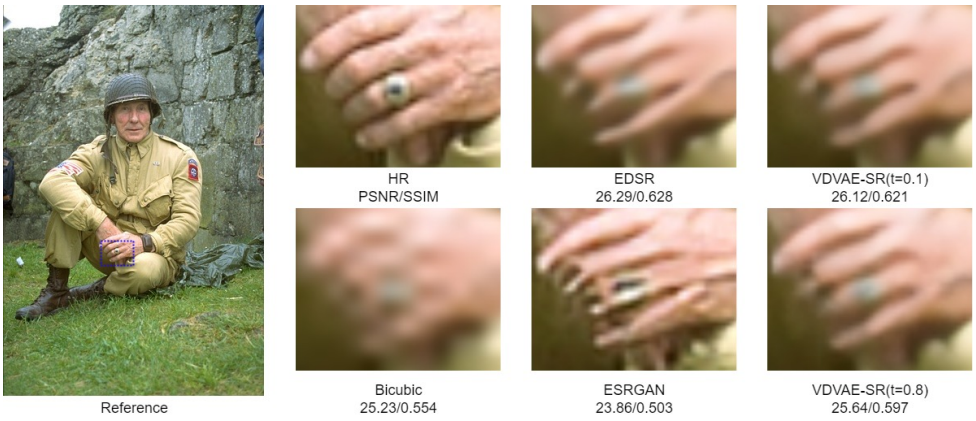

Figure 5: Comparison of super-resolution models for an image (number 38092) of the BSD100 dataset.

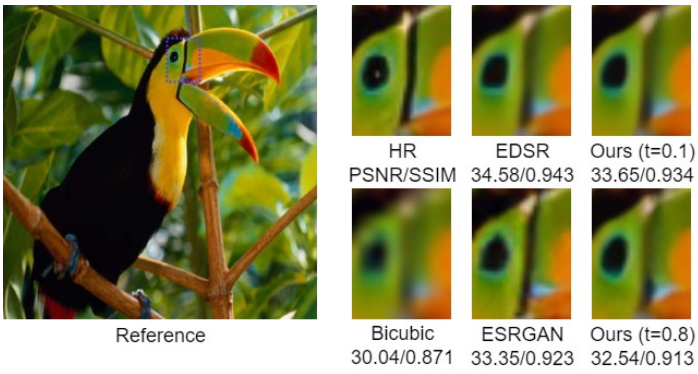

Figure 6: Comparison of super-resolution models for an image of the Set5 dataset.

Figure 7: Prior sampling difference with varying temperature values for 256x256 images (comic image from the Set14 dataset).

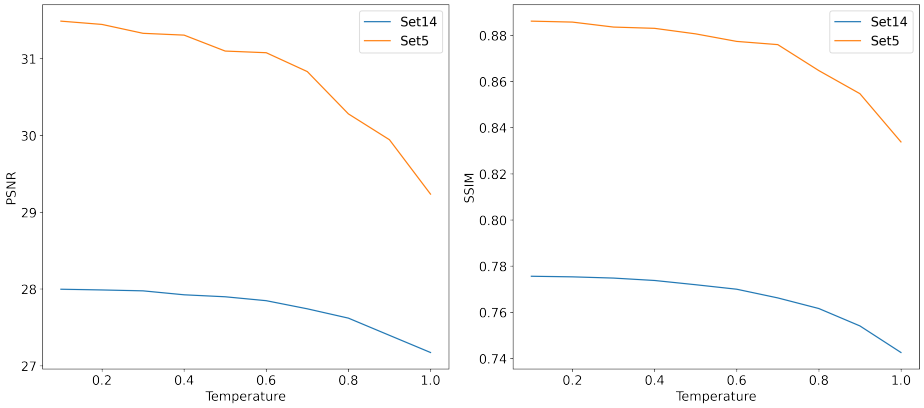

Figure 8: PSNR and SSIM scores of prior samples with varying temperature values for the Set5 and Set14 datasets.

Patch size. A crucial parameter in our super-resolution method is the size of patches to which we apply super-resolution. After experimenting with patches of size 16x16 and 64x64 (i.e., 64x64 and 256x256 after super-resolution), we observed that the 16x16 patch size models were generally performing worse than their counterparts with bigger patch sizes, both in terms of PSNR and SSIM, and in a perceptual sense as the models fail to recreate details that the 64x64 patch models have no problem with. This can also be seen in Fig. 9, especially on the bird's eye, as the general shape and sharpness cannot be recreated by the 16x16 patch size model. We hypothesize that as the patch size gets smaller, the amount of details found in a patch decreases, making the models unable to recreate those details anymore based on context, as the patches will start to look more generic and similar to each other.

6. Conclusions

In this paper, we propose VDVAE-SR, a Very Deep Variational Autoencoder (VDVAE) adapted for the task of image super-resolution (SR). As VDVAEs are expensive to train, we leverage pretrained models and use transfer learning to fine-tune them for super-resolution. We evaluate our method

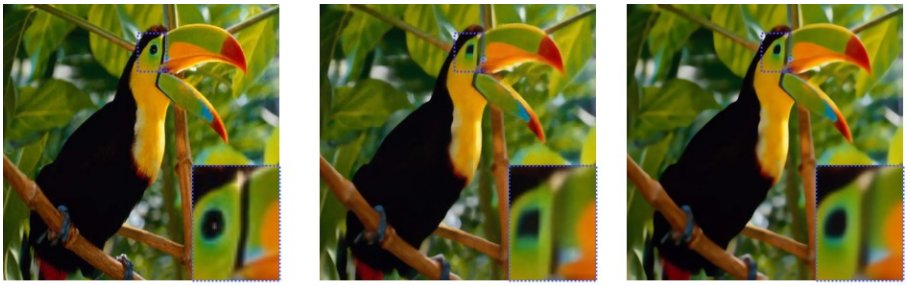

Figure 9: Outputs of models with patch sizes 16x16 and 64x64 on an image from the Set5 dataset.

quantitatively and qualitatively on five common super-resolution datasets and compare it to two popular models, EDSR and ESRGAN. Beside reporting competitive results, we investigate the role of the temperature parameter when sampling. We show that by tuning this parameter at test-time we can control the trade-off between image sharpness and unwanted artifacts that are common in one of our baselines, and achieve a suitable balance between the two. VDVAE-SR is part of the scarce family of VAE-based models for image super-resolution and, to the best of our knowledge, it is the first super-resolution model that employs a very deep hierarchy of latent variables. Thus, we believe that this work demonstrates the promise of this type of approaches, and we hope that it will encourage further research in this underexplored space.

References

Agustsson, Eirikur and Radu Timofte (July 2017). "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Bachlechner, Thomas, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W. Cottrell, and Julian McAuley (2020). *ReZero is All You Need: Fast Convergence at Large Depth*. arXiv: 2003.04887 [cs.LG].

Bevilacqua, Marco, Aline Roumy, Christine Guillemot, and Marie-line Alberi Morel (2012). "Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding". In: *Proceedings of the British Machine Vision Conference*. BMVA Press, pp. 135.1–135.10. ISBN: 1-901725-46-4. DOI: http://dx.doi.org/10.5244/C.26.135.

Brock, Andrew, Jeff Donahue, and Karen Simonyan (2018). "Large scale GAN training for high fidelity natural image synthesis". In: *arXiv preprint arXiv:1809.11096*.

Chen, XI, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel (July 2018). "PixelSNAIL: An Improved Autoregressive Generative Model". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 864–872.

Child, Rewon (2020). "Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images". In: *arXiv preprint arXiv:2011.10650*.

Dai, Tao, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang (June 2019). "Second-Order Attention Network for Single Image Super-Resolution". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dhariwal, Prafulla and Alex Nichol (2021). "Diffusion Models Beat GANs on Image Synthesis". In: *CoRR* abs/2105.05233. arXiv: 2105.05233.

Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang (2015). "Image super-resolution using deep convolutional networks". In: *IEEE transactions on pattern analysis and machine intelligence* 38.2, pp. 295–307.

Gatopoulos, Ioannis, Maarten Stol, and Jakub M Tomczak (2020). "Super-resolution variational auto-encoders". In: *arXiv preprint arXiv:2006.05218*.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). "Generative adversarial nets". In: *Advances in neural information processing systems* 27.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33, pp. 6840–6851.

Ho, Jonathan, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans (2022). "Cascaded diffusion models for high fidelity image generation". In: *Journal of Machine Learning Research* 23.47, pp. 1–33.

Huang, Jia-Bin, Abhishek Singh, and Narendra Ahuja (June 2015). "Single Image Super-Resolution From Transformed Self-Exemplars". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hyun, Sangeek and Jae-Pil Heo (2020). "VarSR: Variational super-resolution network for very low resolution images". In: *European Conference on Computer Vision*. Springer, pp. 431–447.

Jolicoeur-Martineau, Alexia (2018). "The relativistic discriminator: a key element missing from standard GAN". In: *arXiv preprint arXiv:1807.00734*.

Karras, Tero, Samuli Laine, and Timo Aila (2019). "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

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

Kingma, Durk P and Prafulla Dhariwal (2018). "Glow: Generative flow with invertible 1x1 convolutions". In: *Advances in neural information processing systems* 31.

Kingma, Durk P, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling (2016). "Improved variational inference with inverse autoregressive flow". In: *Advances in neural information processing systems* 29.

Ledig, Christian, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. (2017). "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690.

Li, Haoying, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen (2022). "Srdiff: Single image super-resolution with diffusion probabilistic models". In: *Neurocomputing*.

Lim, Bee, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee (2017). "Enhanced deep residual networks for single image super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144.

Maaløe, Lars, Marco Fraccaro, Valentin Liévin, and Ole Winther (2019). "Biva: A very deep hierarchy of latent variables for generative modeling". In: *arXiv preprint arXiv:1902.02102*.

Martin, D., C. Fowlkes, D. Tal, and J. Malik (2001). "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics". In: *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. Vol. 2, 416–423 vol.2.

Matsui, Yusuke, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa (Nov. 2016). "Sketch-based manga retrieval using manga109 dataset". In: *Multimedia Tools and Applications* 76.20, pp. 21811–21838.

Nichol, Alexander Quinn and Prafulla Dhariwal (July 2021). "Improved Denoising Diffusion Probabilistic Models". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8162–8171.

Niu, Ben, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen (2020). "Single image super-resolution via a holistic attention network". In: *European conference on computer vision*. Springer, pp. 191–207.

Oord, Aaron van den, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu koray, Oriol Vinyals, and Alex Graves (2016). "Conditional Image Generation with PixelCNN Decoders". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc.

Oord, Aaron Van, Nal Kalchbrenner, and Koray Kavukcuoglu (June 2016). "Pixel Recurrent Neural Networks". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 1747–1756.

Parmar, Niki, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran (July 2018). "Image Transformer". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 4055–4064.

Pisharoty, Narayan, Manisha Jadhav, and Yogesh Dandawate (Apr. 2013). "Performance Evaluation of Structural Similarity Index Metric in Different Colorspaces for HVS Based Assessment of Quality of Colour Images". In: *International Journal of Engineering and Technology* 5, pp. 1555–1562.

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra (2014). "Stochastic backpropagation and approximate inference in deep generative models". In: *International conference on machine learning*. PMLR, pp. 1278–1286.

Sohn, Kihyuk, Honglak Lee, and Xinchen Yan (2015). "Learning Structured Output Representation using Deep Conditional Generative Models". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc.

Sønderby, Casper Kaae, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther (2016). "Ladder variational autoencoders". In: *arXiv preprint arXiv:1602.02282*.

Timofte, Radu, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, et al. (July 2017). "NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Tong, Tong, Gen Li, Xiejie Liu, and Qinquan Gao (2017). "Image super-resolution using dense skip connections". In: *Proceedings of the IEEE international conference on computer vision*, pp. 4799–4807.

Uria, Benigno, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle (2016). "Neural autoregressive distribution estimation". In: *The Journal of Machine Learning Research* 17.1, pp. 7184–7220.

Vahdat, Arash and Jan Kautz (2020). "Nvae: A deep hierarchical variational autoencoder". In: *arXiv preprint arXiv:2007.03898*.

Wang, Xintao, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy (Sept. 2018). "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks". In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Zeyde, Roman, Michael Elad, and Matan Protter (2010). "On single image scale-up using sparse-representations". In: *International conference on curves and surfaces*. Springer, pp. 711–730.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

Appendix A. Additional results

Here we show additional qualitative test results comparing our VDVAE-SR model with EDSR and ESRGAN on the BSD100 dataset. We show results of our model with sampling temperature of 0.1, 0.8, and 1. Lower temperature reduces the variance of the Gaussian distributions of the prior, resulting in smoother images, but also reducing noise. We believe that sampling with a temperature $t = 0.8$ provides a good trade-off between preserving details and avoiding unnatural artifacts.

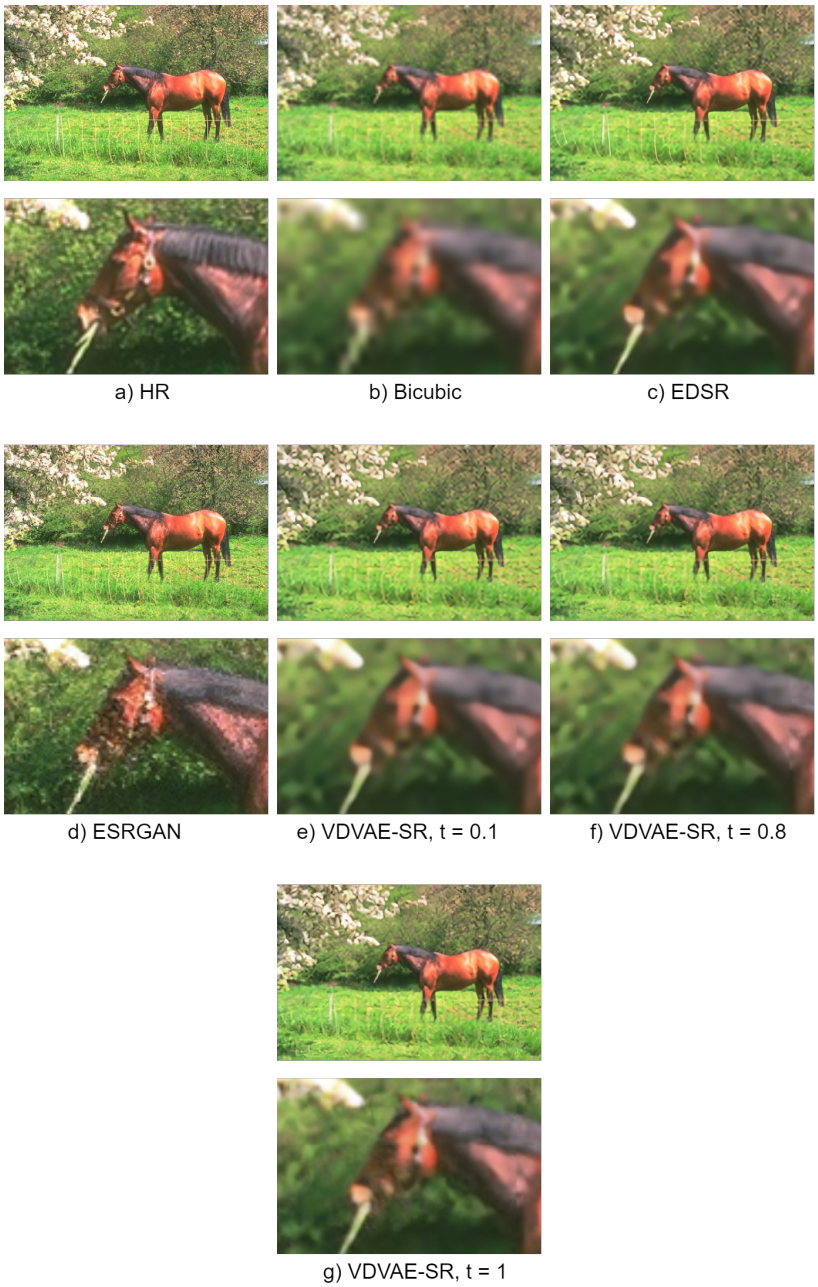

Figure 10: Comparison of super-resolution models for an image (number 291000) from the BSD100 dataset.

Figure 11: Comparison of super-resolution models for an image (number 156065) from the BSD100 dataset.

Figure 12: Comparison of super-resolution models for an image (number 108005) from the BSD100 dataset.

Figure 13: Comparison of super-resolution models for an image (number 157055) from the BSD100 dataset.

Figure 14: Comparison of super-resolution models for an image (number 159008) from the BSD100 dataset.

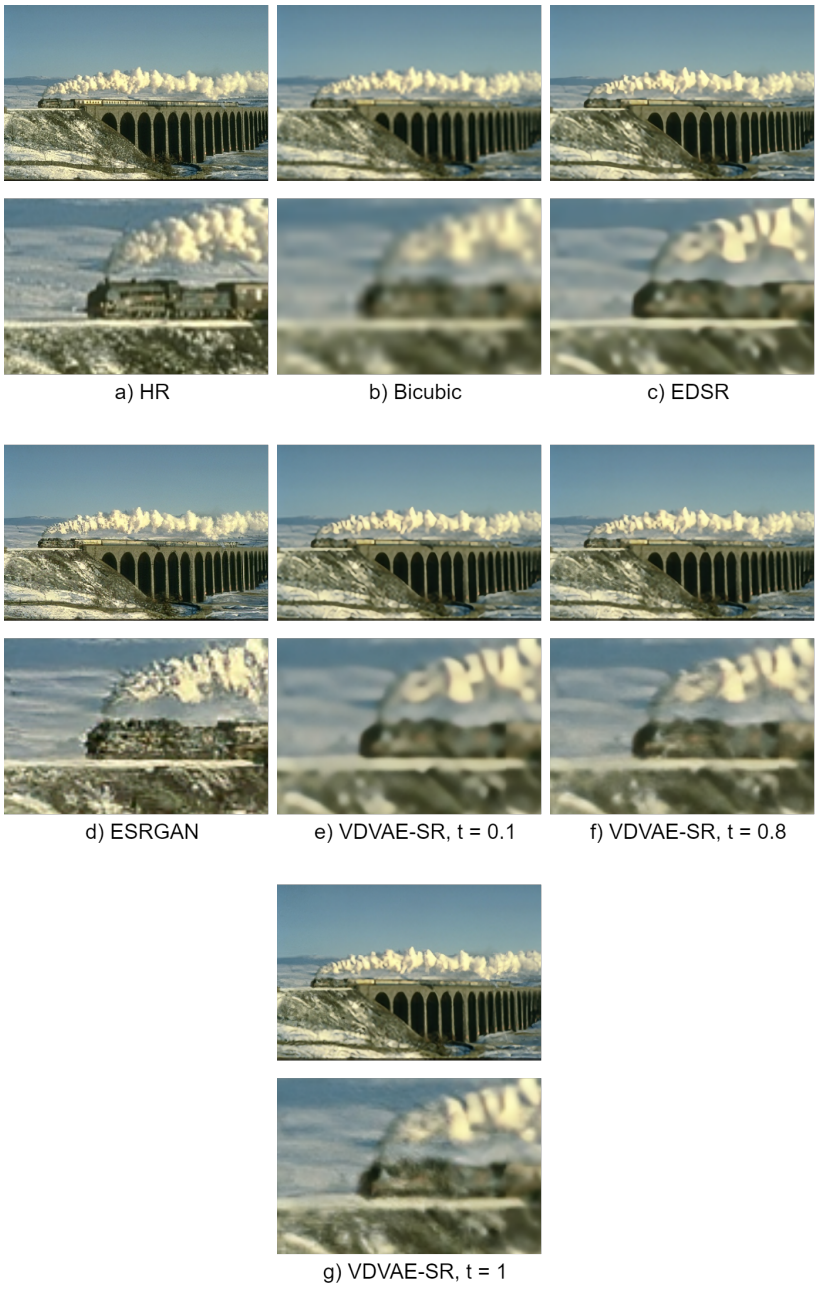

Figure 15: Comparison of super-resolution models for an image (number 182053) from the BSD100 dataset.

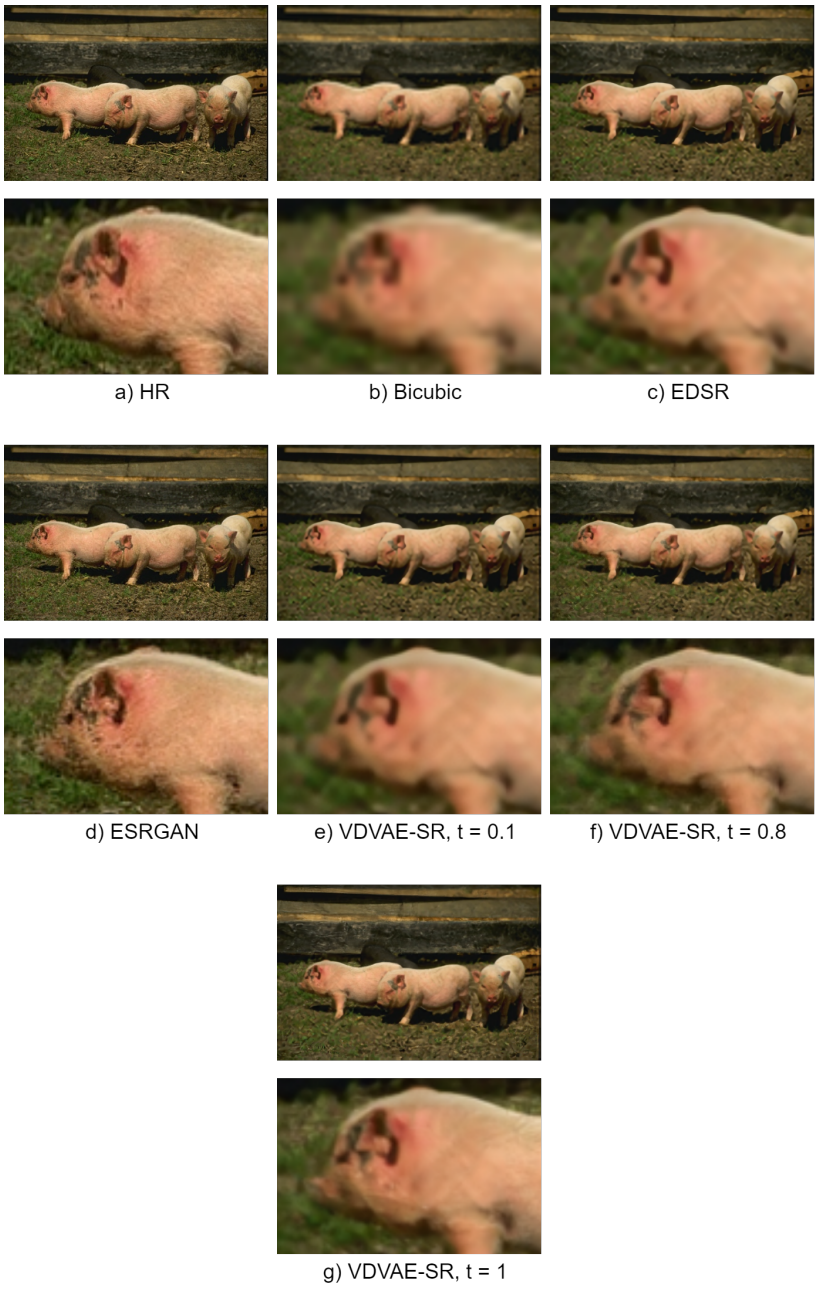

Figure 16: Comparison of super-resolution models for an image (number 66053) from the BSD100 dataset.

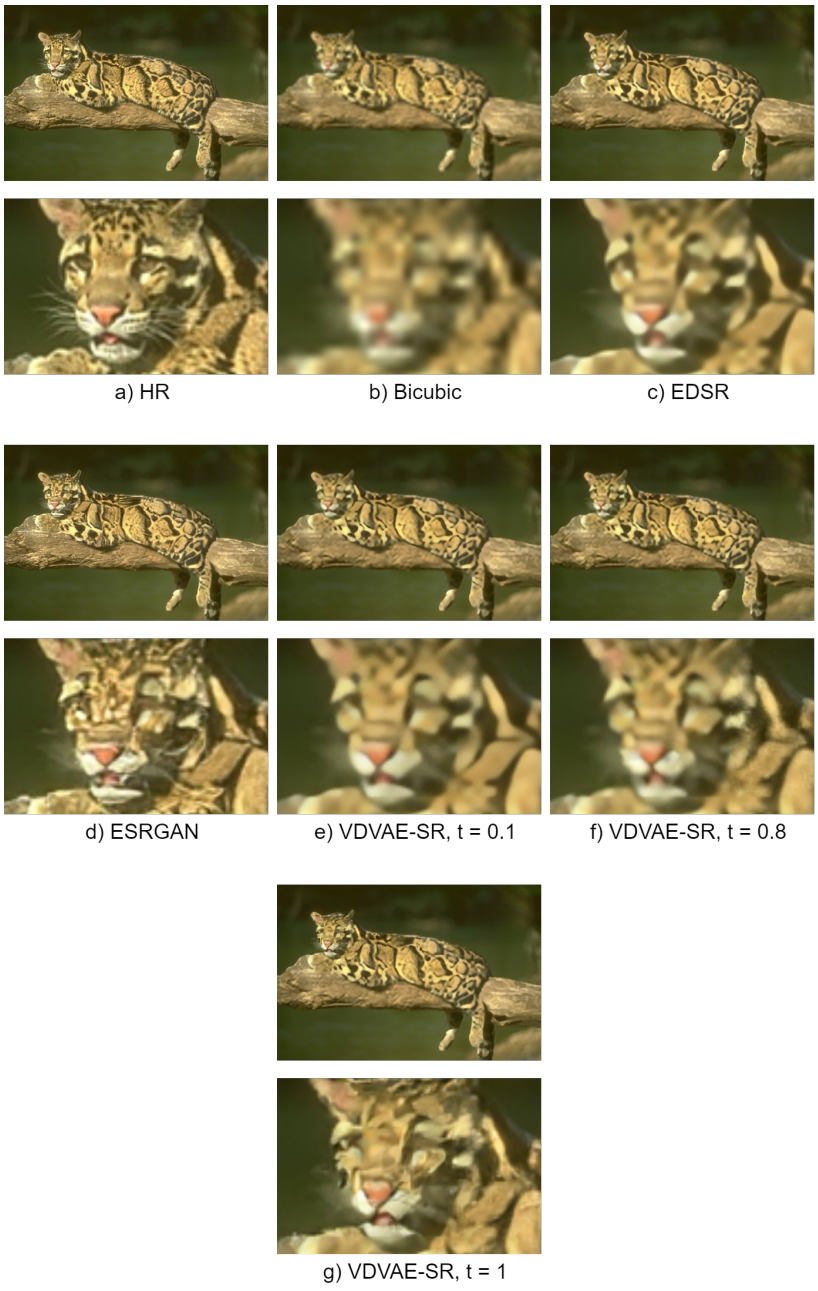

Figure 17: Comparison of super-resolution models for an image (number 160068) from the BSD100 dataset.

Figure 18: Comparison of super-resolution models for an image (number 229036) from the BSD100 dataset.

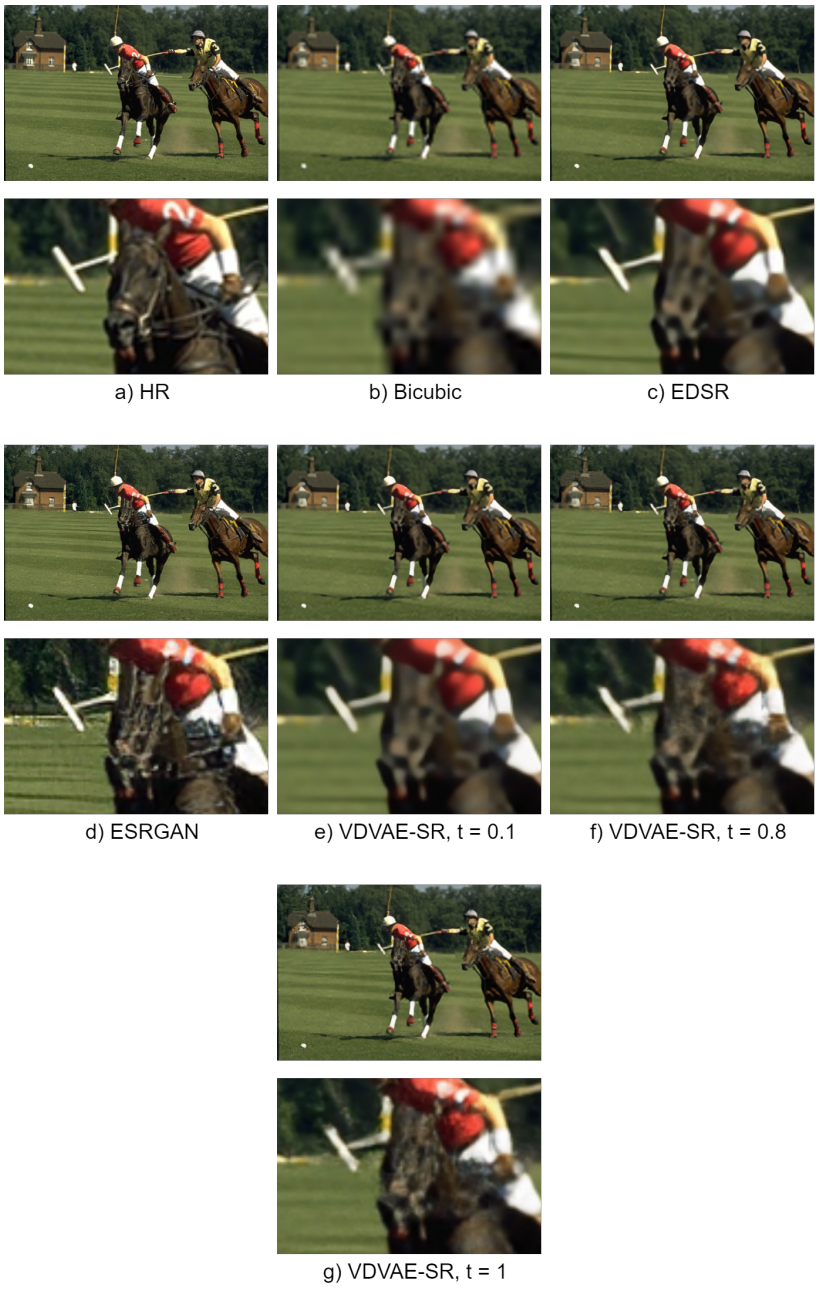

Figure 19: Comparison of super-resolution models for an image (number 361010) from the BSD100 dataset.