We would like to thank the reviewers for their positive feedback, constructive comments and suggestions. Below we respond to the main points raised and will update the paper to take them into account.

Response to Reviewer 1

R1.1: I was missing work on DGM evaluated in feature space, e.g. [1]. Thank you for pointing out this additional reference, we will include it in our discussion of related work. It is indeed related in the sense of using L2 reconstruction losses in abstract feature spaces, rather than in RGB pixel space. They rely on pre-trained ImageNet classification networks to define the feature space. In our work, the feature space is based on invertible transformations, which permits to train the feature space itself in a unified unsupervised learning framework, by optimizing the data likelihood.

R1.2: Is the advantage that the VAE does the generative work and the inverse-layers do not have to be as powerful? Yes, this is precisely the point. The "cheap", but non-invertible, VAE layers are more efficient in the sense that they update all the pixels in every layer, rather than half the pixels in each "coupling layer" of an NVP (each of which consists itself of many CNN layers). This makes the VAE layers mix pixel information more quickly. The VAE can be seen as a "fancy trained prior" for the NVP layers, rather than using a standard normal prior, thus requiring less coupling layers to obtain the desired density model. Equivalently, the NVP layers play the role of a "fancy noise model" that improves over the factored Gaussian over RGB values used in vanilla VAEs. Please see also response R2.3 below.

Response to Reviewer 2

R2.1: Clarity. Thank you for the useful and detailed comments on improving clarity. We will revise the text accordingly, taking these into account one by one. Here we briefly respond to a selection of them for sake of brevity. (1) We will make the link between Sections 3 and 4 more explicit and synthetic. (3) We will improve the captions of figures 1 and 2. (8) BPD is reported in the original papers and was matched exactly when retraining the models ourselves. We will also move the explanation of † and ‡ to the caption of Table 3. (10) We will ensure uniform usage of the terms NVP. (11) More interpretation of the ablation study (Section 5.1) and model refinements (Sec. 5.2 paragraph 1) will be provided.

R2.2: Ablations. – *"[...] So there are more parameters to be trained.":* We would like to first clarify that the invertible transformation f_ψ used in our model is very light. In short, it uses three affine coupling layers, parametrised using two residual blocks rather than eight, implemented as in Real-NVP. In Table 1, the number of weights in the VAE is adjusted to account for the weights in f_ψ, resulting in models with the same number of weights across Table 1. See also line 231 (main paper) and Appendix B. As suggested in Clarity (7), we will make the main text self-contained regarding f_ψ.
– *"[...] A good baseline experiment is the Real-NVP including the adversarial training":* Definitely, in fact the Flow-GAN (ref. [17] in the main paper) trained with hybrid losses provides this baseline. We will emphasize this in Section 5.2 paragraph 2 (please see this section for details). We experimentally compare to Flow-GAN in Table 3 and Figure 6. Results show that we obtain a substantial improvement both in BPD and in sample quality. Note that in our first ablation study (Table 1), simply removing the VAE component results in a tiny flow model (see above), while growing it to be comparable yields the same as the Flow-Gan model.

R2.3: Interpretation. *"Why is your method more efficient[/better] than a Real-NVP with [...] adversarial training?"* The constraint of invertibility in the data space can be costly. Instead, VAEs are able to focus on the low dimensional manifold of natural images, at the cost of approximate inference. Invertible layers used by flow models in practice, affine transformations of half the variables, are quite restrictive (see also response R1.2 above). This is mitigated with expensive residual blocks to parametrize them. From that perspective, using mostly cost-efficient layers from the VAE and a few invertible layers where needed is a cost-efficient choice, which tends to be confirmed by experimental results.

Response to Reviewer 3

R3.1: Would it be possible to improve the compression efficiency of the proposed model? Good question. Without resorting to autoregressive models, which suffer from slow sampling, there are a number of ways to further improve the BPD of our models. (i) The most straightforward is to use "more muscle": training bigger models, and train for more iterations. In the presented experiments, all the models have been trained on a single consumer grade GPUs. (ii) It was recently shown in the Flow++ (ref. [22] in the main paper) that variational dequantization of the discrete pixel values can lead to significant improvements in BPD compared to uniform dequantization. This can also be applied to our model. (iii) Another possibility is to use a discrete flow component, such as [Integer Discrete Flows and Lossless Compression. arXiv:1905.0737] and use a mixture of logistic density functions to model the discrete target features.

R3.2: Would it be feasible to sample from the true distribution for the GAN update? Thanks for raising this question. Yes, this is definitely possible. We experimented with this before submitting the paper, but found the difference in results to be insignificant from using the mean. We will comment on this point in the final version of the paper, and add these results to the supplementary material.