Thank you for your comments. Here we address the most important ones (we can and will also fix the others):

**Rev#1/Rev#4:** What's the difference between Deeploss-VGG/-Squeeze and the loss proposed in [29] (LPIPS)? It is the same metric, as noted in L140. We wanted a consistent naming scheme in the paper, but see that this can be confusing. We consider renaming it to LPIPS-VGG and LPIPS-Squeeze.

**Rev#1:** Comparison to MSE for VAE In our experience there are no big differences on MNIST between $L_2$ and MSE and the SSIM shown (which is in general considered the superior loss for images). If any, the generated images look slightly less crisp, however this effect is often dominated by the quality of hyper-parameter tuning.

Quantitative analysis on other score-functions We will provide a measure in the updated paper. The choice is a bit difficult because, for example, an LPIPS based measure will work best on the models trained with a variant of LPIPS.

**Rev#2:** Evidence that loss actually guides generative models towards better image generations The proposed metric does lead to *much* better results, see Section 4.2, figures 4, 5, and in the supplement Section E, figures E.8–E.11.

Correctness of results in relation to "Deep Feature Consistent Variational Autoencoder" In the article by Hou et al. the weights are not trained but fixed to $\omega_{lc} = 1/C_l$ in our notation. When the weights are adapted based on another dataset, the resulting losses will be different. This is especially true in our case. While the dataset used for tuning is relatively large, it still covers only a small subset of relevant transformations and the number of tunable parameters is large. Thus, overfitting to the dataset can introduce, e.g., warping artefacts.

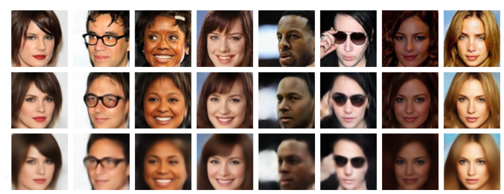

Figure R.1: Reconstructions using VAE. 1st row: Ground truth; 2nd: Watson-DFT, 3rd: CVPR 2019 method (please zoom in).

**Rev#4:** Comparison to "A General and Adaptive Robust Loss Function", Jonathan T. Barron, CVPR, 2019. Thank you, we were not aware of that interesting and relevant publication. There are many difference to our approach: Barron performs a 2D DCT over the entire image while we use a blockwise FFT, which allows us to also consider phase differences. This improves perceptual accuracy significantly compared to DCT. We weight the DCT/FFT frequencies resulting in <200 weights. Barron learns a 'robustness' value for each separate DCT frequency. Due to not using block-DCT, this results in a lot of parameters (e.g., 49152 trainable parameters on 128×128×3 images). We also use the YUV/YCbCr color space, but we weight the color channels by learning the importance of each, while Barron weights them equally. Both approaches learn a "robustness" parameter determining the significance of outliers ($\alpha$ vs. $p$). Barron learns this parameter during training on the generative task, but has to add some regularization to make this work. We learn the perceptual parameters on a perceptual dataset, independent of the generative task and no regularization is necessary. We conducted experiments using the CVPR 2019 method with the code thankfully provided by the author, e.g., see figures R.1 and R.2 (we will add more to the paper, also showing generated images). *The method performs well, but clearly worse than Watson-DFT.*

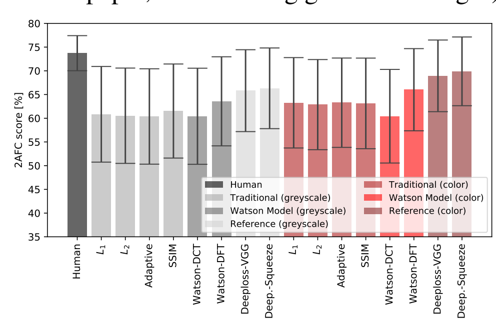


Figure R.2: Updated Figure 6 of the submission, now including $L_1$ and Watson-DCT and the CVPR 2019 method ("Adaptive").

Performance differences in Figure-6 & user-study *This is an important aspect/finding of our study, please revisit L240–L270.* Consider what task is solved and measured in Figure 6: a dataset is generated by applying a certain set of transformations to images. The test-set is not an unbiased estimate of performance in a real application as it does not consider all relevant transformations. Moreover, the generative task differs from the test-set insofar as the VAE, similar to the adversary in a GAN, tends to find the weaknesses in the loss-function in order to maximize the similarity of $q(z|x)$ to $N(0, I)$. A user-study would be nice to have, but may not provide insights into the models. We believe that the differences between the models are so large that there is no need for a user-study to decide which images look better – we assume that all reviewers agree on the obvious visual differences (on random samples, more can be generated using the software provided).

Free parameters & Ablation study of model components A full list of free parameters is given in L137. The maximum number of parameters of our method is 135. Earlier during development we compared DFT and DCT within our model. With DCT the model performed significantly worse in all tasks, on par with the other models. We included this result in Fig. R.2 as "Watson-DCT".

MCMC and the loss as probability distribution We did not want to claim that one can use MCMC if the model is not a valid probability distribution. We will clarify this part. When the loss is a valid unnormalized log-probability, we can use standard MCMC techniques like HMC to sample from $p(x|z)$.