We thank the reviewers for their feedback and are glad that they found the paper to be clear, novel, and a well motivated direction of work. We will incorporate the answers/other feedback into the revised manuscript.

[R1, R2] **The general quality of samples seems to be negatively impacted.** We agree, the qualitative performance of Wavelet Flow (WF) appears to be somewhat worse than the Glow baseline but is still competitive in terms of bits-per-dimension. In particular, global coherence of fine details, eg, eye colour, gaze direction and hair texture, appear inconsistent over larger distances. It is not immediately clear exactly how the wavelet decomposition, receptive field, patch based training, and conditioning, interact to cause this. Initial experiments investigating these factors were inconclusive; however, we suspect that further architecture search, focused specifically on the conditioning networks, may help address these issues. Finally, we agree that quantitative metrics of image quality such as FID would be valuable. While this are not yet standard practice in normalizing flow (NF) papers which rely primarily on log-likelihood and qualitative assessment for evaluation, we will include an FID-based comparison in the final version.

[R3] **Exploration of different wavelet basis** We agree other wavelets could be potentially interesting. Indeed, we believe (see L290) that learning the basis should be possible and promising direction for future work; however as shown, the ability of any orthonormal wavelet basis to provide a natural, spatially coherent scale decomposition of a signal which fits naturally into the NF framework is, we believe, an important contribution in and of itself.

[R2] **Other baseline comparisons for super-resolution (SR).** SR is not claimed as our primary goal/contribution. Rather, it is a fortuitous byproduct of the conditional structure that WF enables. We included the SR results as a demonstration of this capability. A more thorough exploration of WF for SR is a promising direction for future work. We note that a comparison against a conditional structure $p(\mathbf{I}_i|\mathbf{I}_{i-1})$ is difficult because existing SR approaches of this form generate samples $\mathbf{I}_i$ from the full space $\mathbb{R}^{N\times N\times C}$, including regions of space which are inconsistent with the lower resolution image $\mathbf{I}_{i-1}$. In contrast, the wavelet construction means that SR sampling with WF produces images $\mathbf{I}_i$ which are *exactly* consistent with $\mathbf{I}_{i-1}$ in the sense that it is sampling from a lower dimensional manifold. Consequently, direct comparison of likelihood numbers is not possible. Finally, we note that, while training a NF with an overcomplete representation may be possible, it is not obvious how best to do so. Further, we believe that, because the wavelet transform is a bijection, it is a more natural fit with NFs, which are built on bijections.

[R3] **Further comparison of training time.** We recognise that our comparison of training time is imperfect. This was, in part, due to the challenges of training Glow with limited hardware resources. Using the public Glow code, we measured the average number of seconds-per-image over 100 iterations of training on our hardware. We provide timings for LSUN $64 \times 64$ (LS), ImageNet $64 \times 64$ (IN), and CelebA-HQ $64 \times 64$ (CA). Glow yields 0.954, 0.956, and 1.79 seconds-per-image on LS, IS, and CA, resp. WF yields 0.0147, 0.0144, and 0.0291 seconds-per-image on LS, IS, and CA, resp. Together, this is a speedup over over $60\times$. This is enabled, in part, because the smaller, simpler individual conditional models that makeup WF allow for larger batch sizes than is possible with a single monolithic Glow model.

[R3, R4] **Baseline Comparison on CelebA-HQ at** $256 \times 256$**.** Our experiments were focused on affine models on 8-bit data to avoid using separate models for quantitative and qualitative results, also motivated at L209. Note that, in our view, the practice of using a different model (additive vs affine) on different data (5bit vs 8bit) for qualitative vs quantitative evaluation is an unfortunate practice in the field which we sought to avoid. At the reviewers' request we provide quantitative results on CelebA-HQ $256 \times 256$ 5-bit to compare against the value reported in the supplemental material of Glow of 1.03 bpd. Using an additive and an affine WF model with the same setup and training time as the 8-bit model from the paper, we achieve 1.12 and 0.943 bpd, respectively. Finally, we note that higher resolutions do not produce worse results. Table 1 shows that on ImageNet at $64 \times 64$ Wavelet Flow performs better than the baselines. Further, the improvement over the baselines is even more significant than for ImageNet at $32 \times 32$.

[R3] **Value of MCMC with** $T = 0.97$**.** Yes, we believe that MCMC is worth it, even with $T = 0.97$. As shown in Fig. 3, there is a notable improvement in image quality from $T = 1$ to $T = 0.97$. See also Figs. S1-S15 in the Suppl. Mat. which contain more samples with $T = 0.97$ and $T = 1$ (i.e., without annealing). Note we believe the ultimate goal with NF is a model which produces high quality samples without annealing. MCMC allows us to determine how close we are to that goal and the fact that only a mild amount of annealing is required suggests that we are close.

[R3] **Marginal distribution of detail coefficients.** By definition, most detail coefficients are near the peak; capturing that region is likely to have the largest impact on image quality. Because of this we believe that fitting the peak is generally more important than the tails. However, instead of relying entirely on a qualitative comparison of the histograms, we have since computed the KL divergence between the generated histograms and the histogram of ground truth. In this case the annealed affine model (0.0309) matches better than the annealed additive model (0.0365).

[R3] **Difference between sampling from T=1 with and without MCMC?** For the case when $T = 1$ sampling with our MCMC method is effectively equivalent to exact sampling, albeit slower. Because we performed MCMC in the latent space, when $T = 1$ it becomes equivalent to performing MCMC on a Gaussian distribution with identity covariance. In this case, the MCMC method used (the HMC-based, No-U-Turn Sampler) is nearly exact.

[R2] **Intuition about the wavelet coefficients.** At $2 \times 2$, *global variations* capture whether the average intensity of the top/left halves of the image are brighter or darker than the bottom/right halves.

[R4] **Additional implementation details** Training details are in Sec. 3 and (cf L195) specifics of the architecture are in Sec. C of the supp. Code will be released with the final paper (cf L85) and is included with the submission.