High-Fidelity Generative Image Compression – Rebuttal

We thank the reviewers for the positive feedback. In the paper, we extensively study how to combine Generative Adversarial Networks and learned compression to obtain a state-of-the-art generative lossy compression system.

As the reviewers state, our method yields "impressive visual quality" (R3), "high quality reconstructions" (R4), the paper contains "extensive and sound evaluation" (R3) and "extensively studies the proposed architecture, training strategies, as well as the loss" (R4). We show "extremely compelling practical results, and interesting theoretical progress and observations" (R1). The paper is "well written" (R2) and "insightful" (R2). Additionally, "prior work is clearly discussed, critiqued, built upon and it is shown how this work moves to a novel contribution" (R1).

In the following, we address the questions and concerns raised by the reviewers in detail.

R1: Concerns about user study methodology. *"The authors ask users to search for 'interesting' areas of the image [...]. Is it assumed that there are no meaningful compression errors in flat/featureless areas of the image?"* We chose crop size such that the crops are large enough to contain context, while at the same time focusing raters on a part of the image (rating full-resolution images is much harder, as the rater's attention is divided between parts of the image). This results in relatively large (768×768px crops – see Appendix A.8 for a visual example of the study interface) which also contain flat areas. Furthermore, we observe our method actually shines in flat regions (see, e.g., the wall in image CLIC2020/b3f37 in Appendix B). We will release the crop location selected by each user, for each image, for other researchers to compare.

"It's also a bit strange to compare full image ratings from the metrics to crops rated by the individuals." We want to assess whether any metric gets rid of the need to do a user study (L216). When using a metric without a study, user-selected crops would not be available. As to why we use crops and not the full image for raters: it allows us i) to focus the rater's attention (see previous reply) ii) to avoid any downsampling (which would bias results) iii) to show the original next to the reconstruction. Note that each rater selects a different random crop, which together cover most of the image.

"Do you ensure consistency of display and viewing conditions?" While we did not ensure consistency of viewing distance and color space, the study participants were all professionals with access to a well lit office and a modern, sufficiently large, and bright monitor.

R4: Comparison with prior work needed in main text. As we note in L29, Agustsson et al. targeted "extremely low" bitrates and thus operate in a different regime than our models. We nevertheless showed a qualitative comparison in Appendix A.3. There, we also compare to Rippel et al., who did not release models or sufficient images to calculate statistics. We will add a pointer to that Appendix in the main text.

R2: If possible, it would be interesting to compare to VVC. VVC licensing prohibits us from running that code (if we read the licensing terms correctly). We are open to adding a comparison if someone with the proper access rights could run VVC for us (acknowledging them, of course). Since we will open source the model, and the reconstructions, this should be very easy to do for such a person/entity.

R2: What happens when you apply HiFiC to an image sequence? This is an interesting question. Because we focused on images, our method contains no mechanism to guarantee temporal consistency, and inconsistencies in fine detail between frames will likely be visible. In general, we believe that generative video compression is a very interesting direction for future work, and to ensure temporal consistency, mechanisms to enforce smoothness could be borrowed from, e.g., GAN-based video super resolution models. We will mention this in the conclusion.

R3: Discuss failure cases in main text. We will include the relevant part of Appendix B where we discuss the few failure cases ("very small scale text [and] small scale faces" L534–L536) in the main text, and extend it, as part of the additional 9th content page available for the camera ready version. We are interested in seeing how future work addresses these limitations, and will highlight this direction in the conclusion. One could, e.g., include a face detector in the training pipeline, and raise the MSE penalty in the face region, or fall back to other approaches in these regions. In this paper, we focused on improving overall visual quality across a wide variety of images while keeping the method relatively simple.

R4: Why "high-fidelity"? The reviewer is concerned that GANs produce "fake" texture, and finds that the title may thus not be appropriate. We agree that the proposed method may not be pixel-wise accurate to the input, however, the reconstructions are very close semantically and texture-wise. This is validated by the user study (users see the original), as well as the visuals in the main text and Appendix B. We thus think it is justified to call our approach "high-fidelity" in the sense of matching the input image distribution. We also note that "high-fidelity" is commonly used with this meaning in the GAN literature, e.g., in Bock, Donahue, and Simonayan's "Large Scale GAN Training for High Fidelity Natural Image Synthesis" (ICLR 2019, arxiv 1809.11096). We will clarify this perspective in the paper. If the reviewers insist, we could of course change the title, to, e.g., "High-Resolution Generative Image Compression".