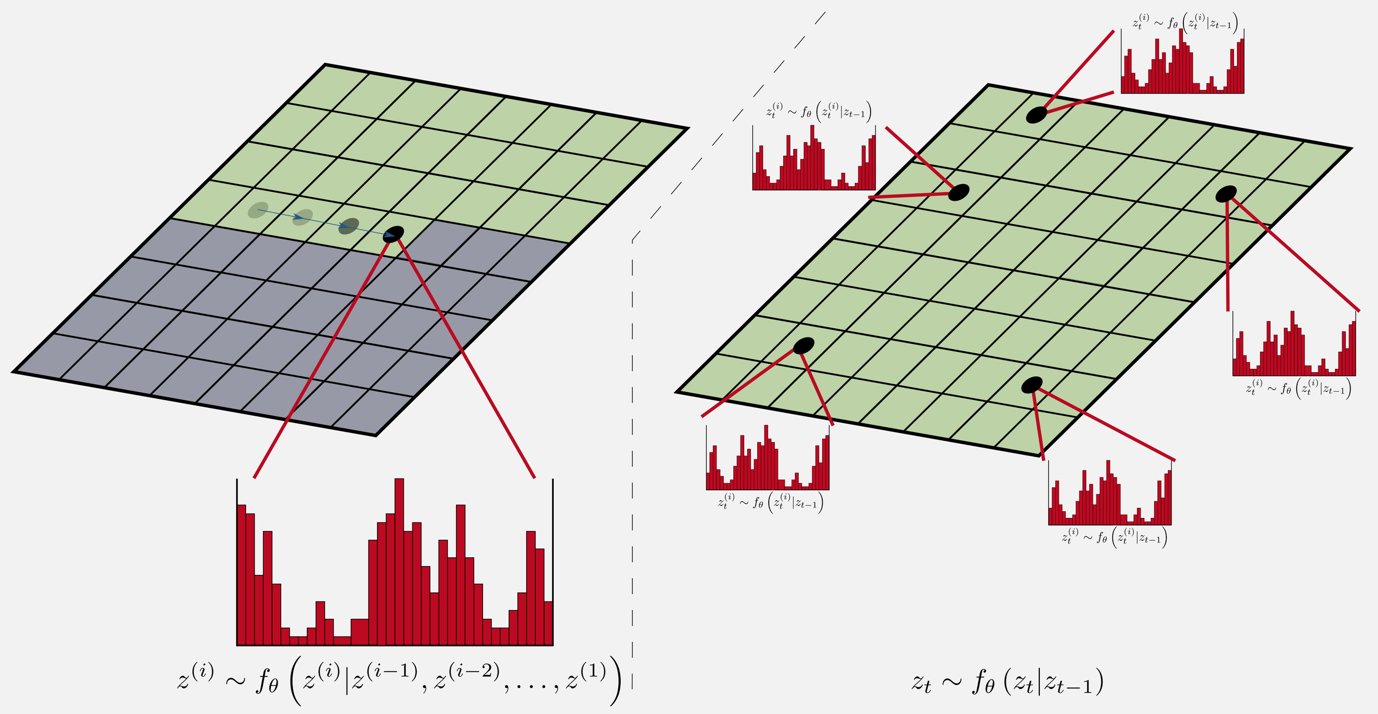
Megapixel Image Generation with Step-unrolled Denoising Autoencoders

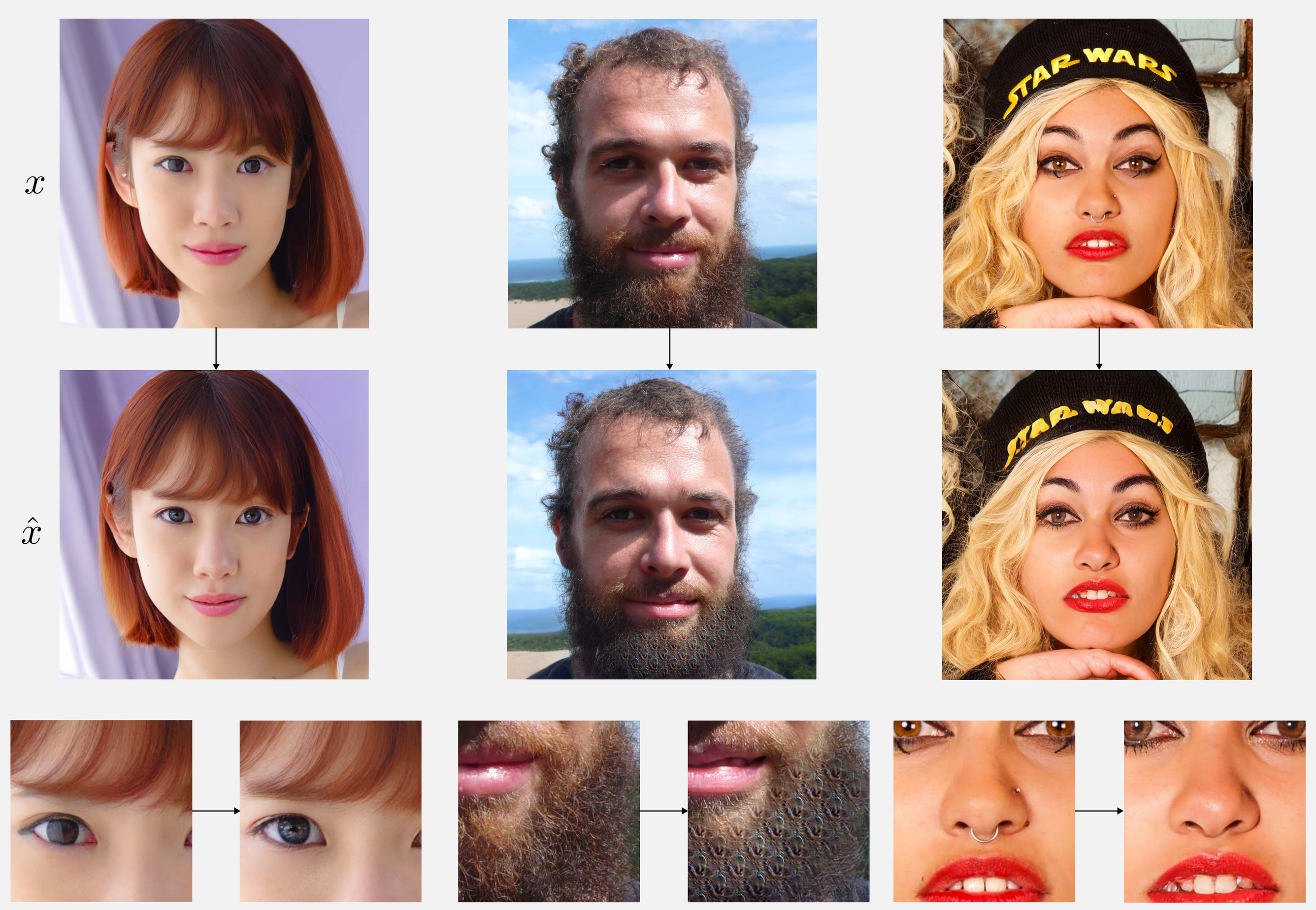
Alex F. McKinney, **Chris G. Willcocks**

Figure 1: Samples from our FFHQ1024 model. Resulting samples are diverse and of high-fidelity. **Each 1024×1024 sample was generated in two seconds** on a consumer-grade GPU (GTX 1080Ti), in contrast to existing approaches at this resolution, which take minutes to generate. To our knowledge, this is the fastest sampling, non-adversarial generative framework at this resolution.

Background

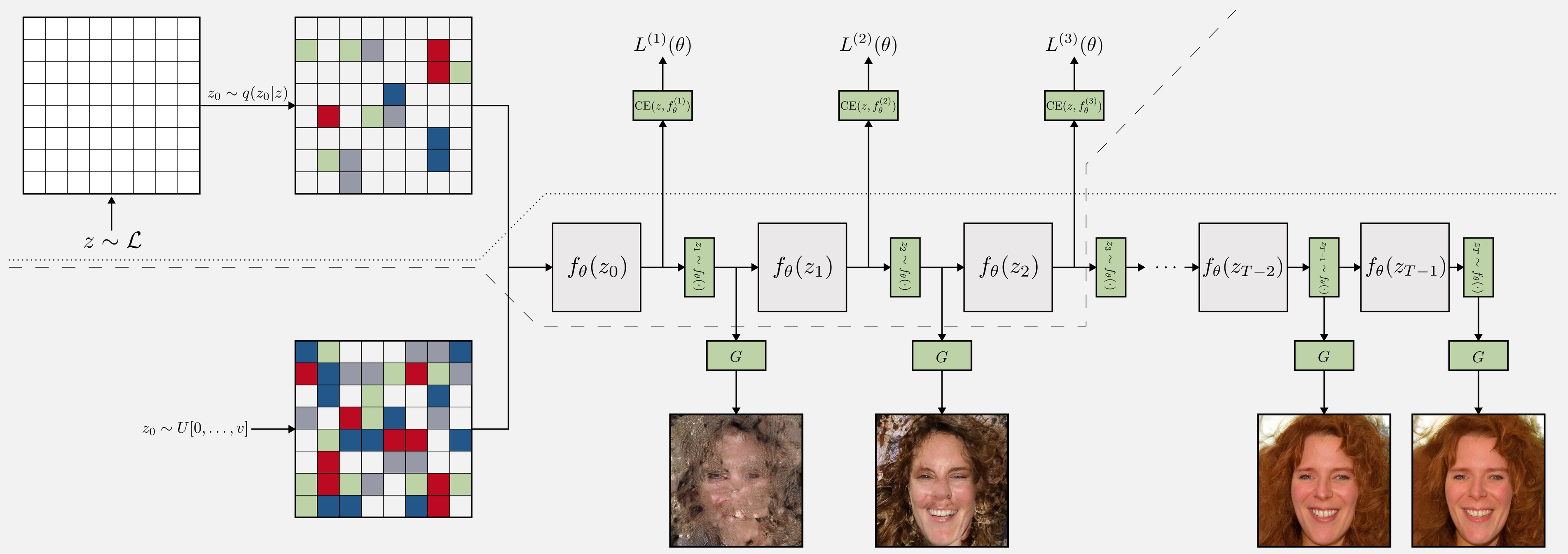


- **Autoregressive (AR)** sampling (left) is defined by the probabilistic chain rule. This means sampling is done iteratively with complexity $\mathcal{O}(n)$.
- **Non-autoregressive (NAR)** sampling (right) samples an arbitrary number of elements in parallel and does not scale with input size n, but may still require thousands of iterations ($\mathcal{O}(1)$ with potentially a large constant).
- AR sampling is **limited to using past context**, whereas NAR uses **all context available to it**.



- **VQ-GAN** is used to reduce computational requirements in generative models by compressing the input into a **discrete space**.
- It offers a higher compression rate than prior work, but does not always faithfully reconstruct the input, for example the left has eye colour changed, and the right has piercings removed.

Proposed Method



An overview of our proposed training and sampling method:

- Above the dashed line shows the training process, beginning by sampling a corrupted sample $\mathbf{z}_0 \sim q(\cdot|\mathbf{z})$. SUNDAE denoises for 2 to 3 steps and averages the cross entropy losses across all steps in the Markov chain.
- Below the dotted line shows the sampling process, beginning by sampling $\mathbf{z}_0$ from a uniform prior distribution. SUNDAE denoises for $T \gg 3$ steps to produce $\mathbf{z}_T$ before using VQ-GAN to decode the final sample.
- To demonstrate the scalability of our approach, we trained a VQ-GAN model from scratch on megapixel images – **considerably larger than prior work**. The loss function is:

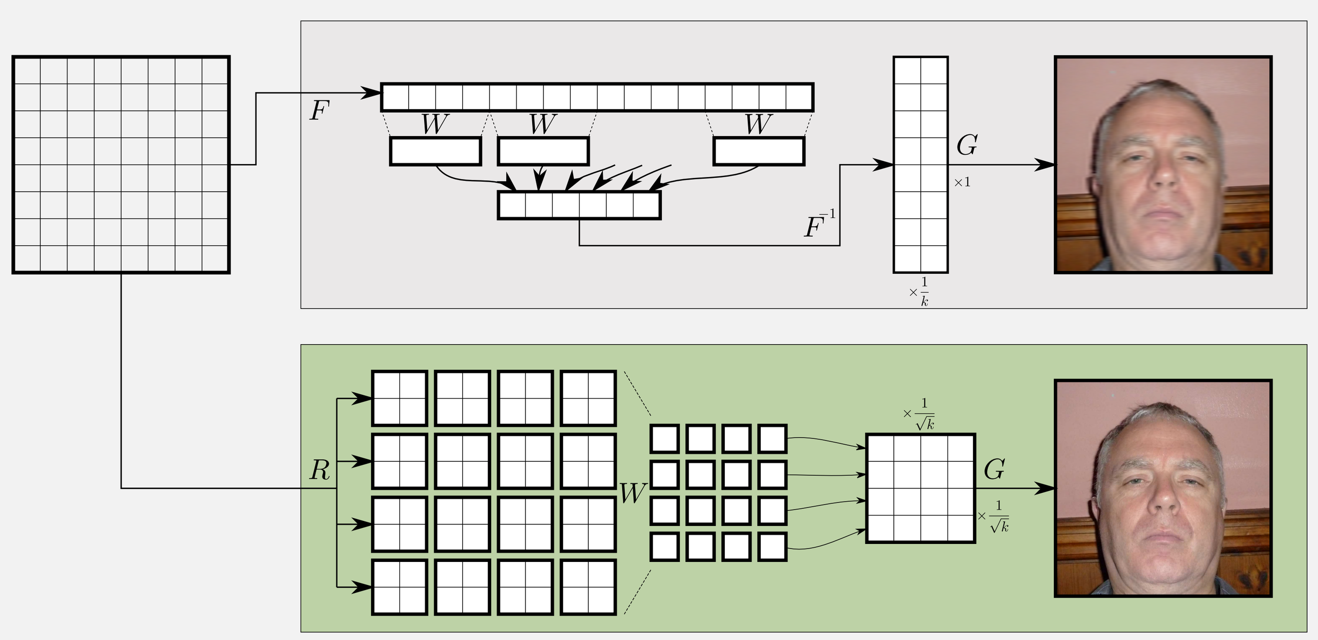
$$
\begin{aligned}
L_{\mathsf{VQ}} &= \alpha_{\mathsf{VQ}} \cdot (||\hat{\mathbf{z}} - \mathbf{z}||^2 \\
&\quad + ||sg[E(\mathbf{x})] - \mathbf{z}||_2^2 + ||E(\mathbf{x}) - sg[\mathbf{z}]||_2^2) \\
L_{\mathsf{PIX}} &= \alpha_{\mathsf{PIX}} \cdot |\mathbf{x} - \hat{\mathbf{x}}| \cdot \\
L_{\mathsf{GAN}} &= \alpha_{\mathsf{GAN}} \cdot (\log D(\mathbf{x}) + \log(1 - D(\hat{\mathbf{x}}))) \\
L &= L_{\mathsf{VQ}} + \lambda \cdot L_{\mathsf{GAN}}
\end{aligned}
$$

$$
\lambda = \frac{\nabla_{G_{-1}}[L_{\mathsf{PIX}} + L_{\mathsf{PER}}]}{\nabla_{G_{-1}}[L_{\mathsf{GAN}}] + \epsilon}
$$

$$
\alpha_{\mathsf{PIX}} = 1.0,\ \alpha_{\mathsf{VQ}} = 1.0,
$$

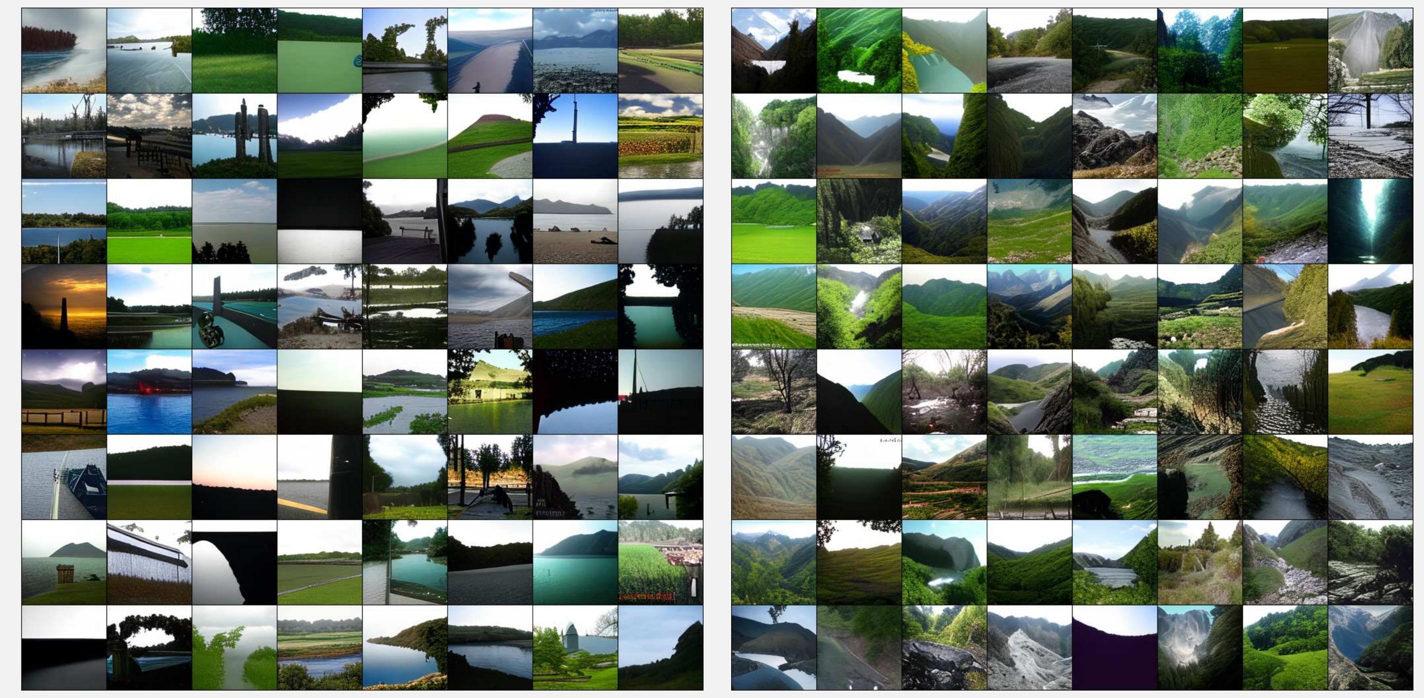
$$
\alpha_{\mathsf{GAN}} = 0.5,\ \alpha_{\mathsf{PER}} = 1.0
$$

- By applying our framework to discrete latent representations, we obtained a fast and scaleable generative model on megapixel images.
- Our method generates 1024×1024 samples in only **two seconds** – a wide margin faster than prior non-adversarial methods which take **minutes** to generate at this resolution.



- As a result of our work, we also found flaws in the original formulation of hourglass transformers when applied to multi-dimensional data.
- Our modifications have **applications in a wider context**, outside of generative modelling.

Results



- 256×256 **class-conditioned** samples from our ImageNet model.
- Left and right batches are from classes “Lakeside” and “Valley” respectively.
- Our approach is easily extendable to use text prompts, **yielding a fast text-to-image generator**.

- Example inpainting results using our FFHQ1024 model. We show multiple results given the same image-mask pair to **demonstrate diversity in the outputs**.
- NAR methods allow for **arbitrary inpainting patterns** to be easily used. In contrast, AR models cannot easily handle all patterns, nor can they use all context when inpainting.

References

[1] Patrick Esser, Robin Rombach, and Björn Ommer. *Taming Transformers for High-Resolution Image Synthesis*. 2021. arXiv: 2012.09841 [cs.CV].

[2] Piotr Nawrot et al. *Hierarchical Transformers Are More Efficient Language Models*. 2021. arXiv: 2110.13711 [cs.LG].

[3] Nikolay Savinov et al. *Step-unrolled Denoising Autoencoders for Text Generation*. 2022. arXiv: 2112.06749 [cs.CL].