Rejection Sampling IMLE: Designing Priors for Better Few-Shot Image Synthesis

Chirag Vashist, Shichong Peng, and Ke Li

APEX Lab
School of Computing Science
Simon Fraser University
{chirag_vashist, shichong_peng, keli}@sfu.ca

A Theory

In this section, we try to justify our

$$m\left(1-F_{\tilde{D}_{i1}}(t)\right)^{m-1} f_{\tilde{D}_{i1}}(t) = n\left(1-F_{D_{i1}}(t)\right)^{n-1} f_{D_{i1}}(t)$$
$$\implies f_{\tilde{D}_{i1}}(t) = \frac{n}{m}\frac{\left(1-F_{D_{i1}}(t)\right)^{n-1}}{\left(1-F_{\tilde{D}_{i1}}(t)\right)^{m-1}} f_{D_{i1}}(t) \tag{1}$$

Notice that because of $\left(1-F_{\tilde{D}_{i1}}(t)\right)^{m-1}$ term in the denominator, we have to make sure that the expression for $f_{\tilde{D}_{i1}}(t)$ is bounded. One way to do that is to truncate the right tail of the ideal distribution $f_{D_{i1}}(t)$ to 0. More explicitly, for a very large T (e.g., $T = 100,000$), we can write:

$$g_{D_{i1}}(t) = \begin{cases} f_{D_{i1}}(t) & \text{if } t \leq T \\ 0 & \text{if } t > T \end{cases}$$

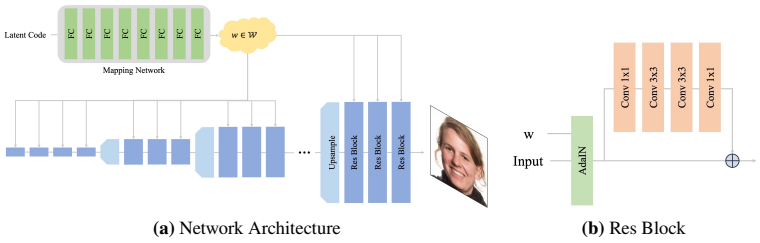
Here $g_{D_{i1}}(t)$ is the PDF of the truncated distribution. Since very large values of distances (t) are rarely observed at test time, so applying this truncation has little effect in practice. Instead of writing the expression for Equation 1 in terms of $g_{D_{i1}}(t)$, we continue to use $f_{D_{i1}}(t)$ along with a constant c associated with the truncation.

Hence using $\phi(t) = \frac{n}{m}\frac{\left(1-F_{D_{i1}}(t)\right)^{n-1}}{\left(1-F_{\tilde{D}_{i1}}(t)\right)^{m-1}}$ and c as the constant associated with the truncation described above, we can write Equation 1 as:

$$f_{\tilde{D}_{i1}}(t) = c\phi(t) f_{D_{i1}}(t) \tag{2}$$

B Network Architecture

Our network architecture is illustrated in Figure 1, comprising a fully-connected mapping network inspired by [2] and a generator network constructed using decoder modules from VDVAE [1]. We choose an input latent dimension of 1024 for all datasets.



(a) Network Architecture **(b)** Res Block

Fig. 1: (a) Network architecture, which comprises of a mapping network, upsampling layers and res blocks (details in (b)). (b) Inner workings of res blocks.

C Experiments

Table-1 gives the details about the number of images in each dataset as well as the value of radius used in the rejection sampling procedure (epsilon, ϵ) used in the results presented in the main paper. The selection of epsilon values was conducted through the process of hyperparameter tuning. We present an ablation study with different values of epsilon later in the paper.

	Obama	Grumpy Cat	Panda	FFHQ-100	Cat	Dog	Anime	Skulls	Shells
Num. of Images	100	100	100	100	160	389	120	96	64
Epsilon Used	0.15	0.18	0.18	0.15	0.15	0.15	0.18	0.18	0.18

Table 1: Number of images in each dataset and the value of epsilon used.

C.1 Random samples

In Figure 2, we compare the random samples of our method to that of the baseline for more datasets.

C.2 Visual Recall

Figure 3 shows the results for the proposed Visual Recall test for more queries. Note how the images produced by our method are the closest to the query and yet have diverse *meaningful* changes.

Since the images displayed are the *nearest neighbours* of the query images, it would be valuable to emphasize the subtle distinctions in the samples produced by our method. In Figure 3a and 3b, we can notice a change in the texture and color of the skin and hair of our samples. In Figure 3c and 3d, we can observe subtle changes to the jaw structure, number of teeth and hue of the different skull samples. Similarly in Figure 3e, we can notice subtle changes in the color of the fur and tilt of the head for different cat samples. In Figure 3g, we observe diversity in hair color, background and ear of the produce samples.

C.3 Ablation on latent dimensions and model parameters

Method	Dim.	Params.	Anime	Shells	Skulls
FastGAN	256	29M	69.8	120.9	109.6
FakeCLR	512	24M	77.7	148.4	106.5
FreGAN	256	147M	59.8	169.3	163.3
ReGAN	512	24M	110.8	236.1	130.7
AdaIMLE	1024	36M	65.8	108.5	81.9
RS-IMLE	1024	36M	35.8	55.4	51.1
	512	19M	48.5	52.9	60.1
	256	12M	53.8	71.7	64.3

Table 2: Comparison between different methods: latent dimensions and number of trainable parameters. Last three columns are FID on Anime, Shells and Skulls dataset.

Table 2 gives the details about the architectures used by the different methods. To decouple the impact of our proposed method (RS-IMLE) from architectural choices, we train using our method using lower latent dimensions. At lower dimensions, the number of parameters for RS-IMLE are significantly lower compared to the other methods. We tabulate the FID for the three most challenging datasets in the last three columns of Table 2. As we decrease the number of dimensions (and consequently the number of parameters), we observe a slight drop in the FID for our method. However, even at *significantly* lower parameter count, our method *outperforms* the baselines.

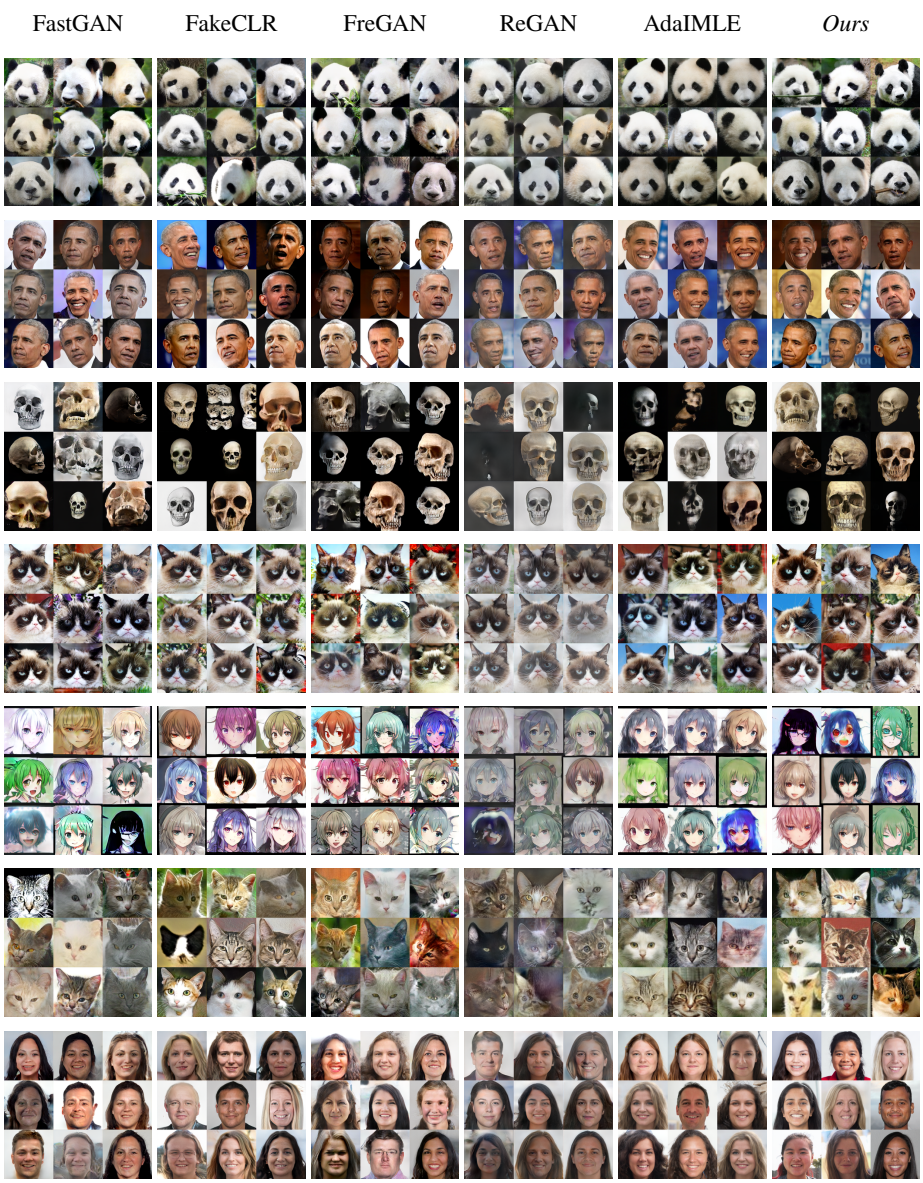


Fig. 2: Qualitative comparison between our method and baselines. While analyzing the images, look for the sharpness of each image and diversity in the content of all images for a method.

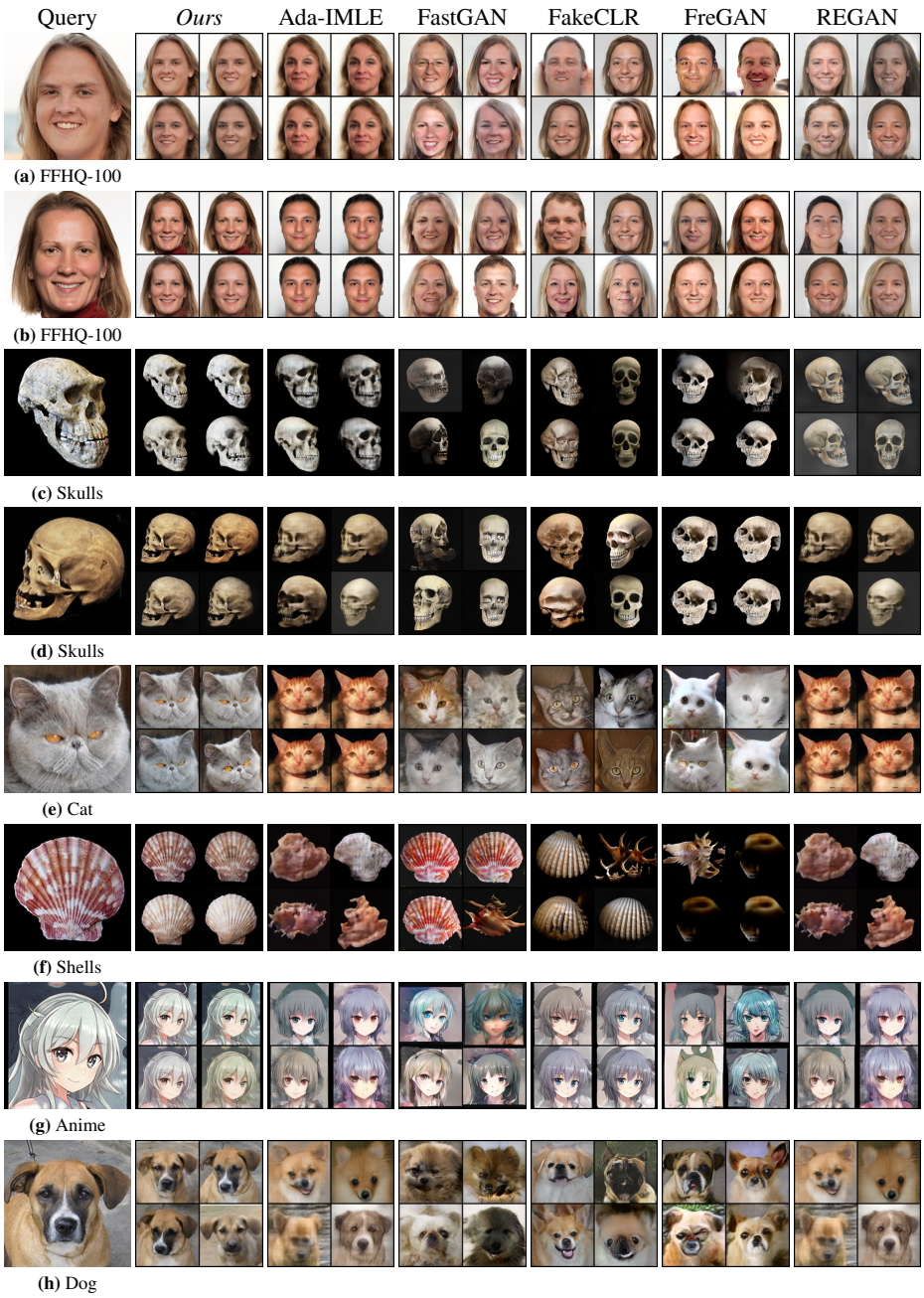


Fig. 3: Visual Recall Test: First column is the query image from the dataset. Subsequent columns are the samples produced by different methods that are closest to the query image in LPIPS feature space. The samples produced by our method are closer to the query images compared to the baselines, while being sufficiently diverse.

References

1. Child, R.: Very deep vaes generalize autoregressive models and can outperform them on images. ArXiv **abs/2011.10650** (2021)
2. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)