Normalizing Flows with Multi-Scale Autoregressive Priors

Apratim Bhattacharyya* [1] **Shweta Mahajan*** [2] **Mario Fritz** [3] **Bernt Schiele** [1] **Stefan Roth** [2]

Abstract

Split coupling normalizing flow-based models admit efficient inference and sampling for image synthesis. However, owing to the efficiency constraints, they have limited expressiveness for modeling long-range data dependencies compared to autoregressive models which rely on conditional pixel-wise generation. In this work, we improve the representational power of split coupling flow-based models by introducing channel-wise dependencies in their latent space through multi-scale autoregressive priors (mAR). Our mAR prior for models with split coupling flow layers (mAR-SCF) can better capture dependencies in complex multimodal data. The resulting model not only achieves state-of-the-art density estimation results on CIFAR-10 and ImageNet but also allows for improved image generation quality, with gains in FID and Inception scores compared to state-of-the-art flow-based models.

1. Introduction

Autoregressive models (Domke et al., 2008; van den Oord et al., 2016a;b) and normalizing flow-based generative models (Dinh et al., 2015; 2017; Kingma & Dhariwal, 2018) are two classes of exact inference models successful on complex high dimensional data e.g. images. Autoregressive models capture complex and long-range dependencies between the dimensions of a distribution. However, the main limitation is that sampling is sequential and thus difficult to parallelize. Normalizing flow-based models, allow exact inference by mapping the input data to a known base distribution, e.g. a Gaussian, through a series of invertible transformations. These models leverage invertible split coupling flow (SCF) layers in which certain dimensions are left unchanged by the invertible transformation as well as SPLIT operations following which certain dimensions do not undergo subsequent transformations. This allows for considerably easier parallelization of both inference and generation. However, their performance still lags behind autoregressive approaches.

In this work, we *(i)* propose multi-scale autoregressive priors for invertible flow models with split coupling flow layers, termed *mAR-SCF*, to address the limited modeling power of non-autoregressive invertible flow models (Dinh et al., 2017; Ho et al., 2019; Kingma & Dhariwal, 2018; Razavi et al., 2019); *(ii)* we apply our multi-scale autoregressive prior after every SPLIT operation such that the computational cost of sampling grows linearly in the spatial dimensions of the image compared to the quadratic cost of traditional autoregressive models (given sufficient parallel resources); *(iii)* our experiments show that we achieve state-of-the-art density estimation results on CIFAR10 (Krizhevsky et al., 2009), and ImageNet (Russakovsky et al., 2015) compared to prior invertible flow-based approaches; and finally *(iv)* we show that our multi-scale autoregressive prior leads to better sample quality as measured by the FID metric (Heusel et al., 2017) and the Inception score (Salimans et al., 2016), significantly lowering the gap to GAN approaches (Radford et al., 2016; Wei et al., 2018).

2. Overview and Background

We begin with an overview of normalizing flow based models. Specifically, normalizing flows consist of a sequence of n invertible functions f_{θ_i}, which transform a density on the data $\mathbf{x}$ to a density on latent variables $\mathbf{z}$,

$$\mathbf{x} \overset{f_{\theta_1}}{\longleftrightarrow} \mathbf{h}_1 \overset{f_{\theta_2}}{\longleftrightarrow} \mathbf{h}_2 \cdots \overset{f_{\theta_n}}{\longleftrightarrow} \mathbf{z}. \tag{1}$$

Given that we can compute the likelihood of $p(\mathbf{z})$, the likelihood of the data $\mathbf{x}$ under the transformation f can be computed using the change of variables formula,

$$\log p_\theta(\mathbf{x}) = \log p(\mathbf{z}) + \sum_{i=1}^{n} \log \left|\det J_{\theta_i}\right|, \tag{2}$$

where $J_{\theta_i} = {}^{\partial \mathbf{h}_i}/_{\partial \mathbf{h}_{i-1}}$ is the Jacobian of the invertible transformation f_{θ_i} going from $\mathbf{h}_{i-1}$ to $\mathbf{h}_i$ with $\mathbf{h}_0 \equiv \mathbf{x}$. Prior work (Chen et al., 2019; Dinh et al., 2015; 2017; Ho et al.,

*Equal contribution [1]Max Planck Institute for Informatics, Saarland Informatics Campus [2]Department of Computer Science, TU Darmstadt [3]CISPA Helmholtz Center for Information Security, Saarland Informatics Campus. Correspondence to: Apratim Bhattacharyya <abhattac@mpi-inf.mpg.de>.

Second workshop on *Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models* (ICML 2020), Virtual Conference

2019; Kingma & Dhariwal, 2018) considers i.i.d Gaussian likelihood models of $\mathbf{z}$, e.g. $p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \,|\, \mu, \sigma)$.

These models, however, have limitations. First, the requirement of invertibility constrains the class of functions f_{θ_i} to be monotonically increasing (or decreasing), thus limiting expressiveness. Second, of the three possible variants of f_{θ_i} to date (Ziegler & Rush, 2019), MAF (masked autoregressive flows), IAF (inverse autoregressive flows), and SCF (split coupling flows), MAFs are difficult to parallelize due to sequential dependencies between dimensions and IAFs do not perform well in practice. SCFs strike the right balance with respect to parallelization and modeling power. In detail, SCFs partition the dimensions into two equal halves *and* transform one of the halves $\mathbf{r}_i$ conditioned on $\mathbf{l}_i$, leaving $\mathbf{l}_i$ unchanged and thus not introducing any sequential dependencies (making parallelization easier). Example include the affine couplings of RealNVP (Dinh et al., 2017) and MixLogCDF couplings of Flow++ (Ho et al., 2019).

In practice, SCFs are organized into blocks (Dinh et al., 2017; Kingma & Dhariwal, 2018) to maximize efficiency such that each f_{θ_i} typically consists of SQUEEZE, STEPOFFLOW, and SPLIT operations. SQUEEZE trades off spatial resolution for channel depth. For an intermediate layer $\mathbf{h}_i$ of size $[\mathrm{C}_i, \mathrm{N}_i, \mathrm{N}_i]$, the SQUEEZE operation transforms it into size $[4\,\mathrm{C}_i, \mathrm{N}_i/2, \mathrm{N}_i/2]$ by reshaping 2×2 neighborhoods into 4 channels. STEPOFFLOW is a series of SCF (several) coupling layers and invertible 1×1 convolutions.

The SPLIT operation (distinct from split couplings) splits an intermediate layer $\mathbf{h}_i$ into two halves $\{\mathbf{l}_i, \mathbf{r}_i\}$ of size $[2\,\mathrm{C}_i, \mathrm{N}_i/2, \mathrm{N}_i/2]$ each. Subsequent invertible layers $f_{\theta j>i}$ operate only on $\mathbf{r}_i$, leaving $\mathbf{l}_i$ unchanged. That is, the SPLIT operation fixes some dimensions of the latent representation $\mathbf{z}$ to $\mathbf{l}_i$ as they are not transformed any further. This leads to a significant reduction in the amount of computation and memory needed. In the following, we denote the spatial resolutions at the n different levels as $\mathrm{N} = \{\mathrm{N}_0, \cdots, \mathrm{N}_n\}$, with $N = \mathrm{N}_0$ being the input resolution. Similarly, $\mathrm{C} = \{\mathrm{C}_0, \cdots, \mathrm{C}_n\}$ denotes the number of feature channels, with $C = \mathrm{C}_0$ being the number of input channels.

In practice, due to limited modeling flexibility, prior SCF-based models (Dinh et al., 2017; Ho et al., 2019; Kingma & Dhariwal, 2018) require many SCF coupling layers in f_{θ_i} to model complex distributions, e.g. images. This in turn leads to high memory costs and less efficient sampling.

3. Multi-scale Autoregressive Flow Priors

We propose to leverage the strengths of autoregressive models to improve split coupling normalizing flow models e.g. (Kingma & Dhariwal, 2018). Specifically, we propose novel *multi-scale autoregressive priors for split coupling flows (mAR-SCF)*. Using them allows us to learn complex

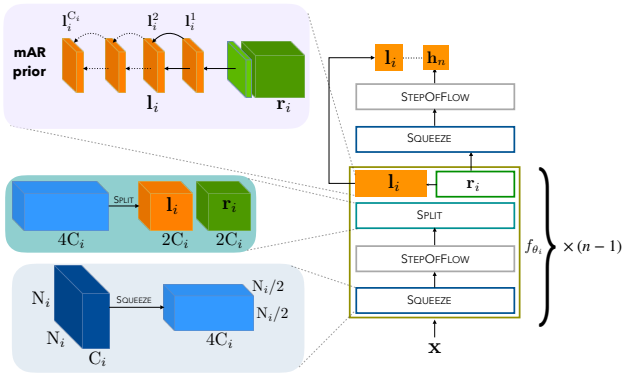


Figure 1. Flow-based generative models with our multi-scale autoregressive prior (*mAR-SCF*). Our *mAR* prior is applied along the channels of $\mathbf{l}_i$, i.e. at each level i after the SPLIT operation.

multimodal latent priors $p(\mathbf{z})$ in multi-scale SCF models, *cf.* Eq. (2) and Fig. 1. This is unlike (Dinh et al., 2017; Ho et al., 2019; Kingma & Dhariwal, 2018; Razavi et al., 2019), which rely on Gaussian priors in the latent space.

Our *mAR-SCF* model (Fig. 1) uses an efficient invertible split coupling flow $f_{\theta_i}(\mathbf{x})$ to map the distribution over the data $\mathbf{x}$ to a latent variable $\mathbf{z}$ and then models an autoregressive *mAR* prior $p_\phi(\mathbf{z})$ (Fig. 1 *top left*), parameterized by ϕ. The likelihood of a data point $\mathbf{x}$ of dimensionality $[C, N, N]$ can be expressed as

$$\log p_{\theta,\phi}(\mathbf{x}) = \log p_\phi(\mathbf{z}) + \sum_{i=1}^{n} \log \left|\det J_{\theta_i}\right|. \qquad (3)$$

Here, J_{θ_i} is the Jacobian of the invertible transformations f_{θ_i}. Note that, as $f_{\theta_i}(\mathbf{x})$ is an invertible function, $\mathbf{z}$ has the same total dimensionality as the input data point $\mathbf{x}$.

Formulation of the *mAR* prior. We now introduce our *mAR* prior $p_\phi(\mathbf{z})$ along with our *mAR-SCF* model, which combines the split coupling flows f_{θ_i} with an *mAR* prior. As shown in Fig. 1, our *mAR* prior is applied after every SPLIT operation of the invertible flow layers as well as at the smallest spatial resolution. Let $\mathbf{l}_i = \{\mathbf{l}_i^1, \cdots, \mathbf{l}_i^{\mathrm{C}_i}\}$ be the C_i channels of size $[\mathrm{C}_i, \mathrm{N}_i, \mathrm{N}_i]$, which do not undergo further transformation f_{θ_i} after the SPLIT at level i. Following the SPLIT at level i, our *mAR* prior is modeled as a conditional distribution, $p_\phi(\mathbf{l}_i|\mathbf{r}_i)$; at the coarsest spatial resolution it is an unconditional distribution, $p_\phi(\mathbf{h}_n)$. Thereby, we assume that our *mAR* prior at each level i autoregressively factorizes along the channel dimension as

$$p_\phi(\mathbf{l}_i|\mathbf{r}_i) = \prod_{j=1}^{\mathrm{C}_i} p_\phi\left(\mathbf{l}_i^j \middle| \mathbf{l}_i^1, \cdots, \mathbf{l}_i^{j-1}, \mathbf{r}_i\right). \qquad (4)$$

Furthermore, the distribution at each spatial location (m, n)

within a channel $\mathbf{l}_i^j$ is modeled as a conditional Gaussian,

$$p_\phi(l^j_{i(m,n)}|\mathbf{l}_i^1, \cdots, \mathbf{l}_i^{j-1}, \mathbf{r}_i) = \mathcal{N}\Big(\mu^j_{i(m,n)}, \sigma^j_{i(m,n)}\Big). \quad (5)$$

Thus, the mean, $\mu^j_{i(m,n)}$ and variance, $\sigma^j_{i(m,n)}$ at each spatial location are autoregressively modeled along the channels. This allows the distribution at each spatial location to be highly flexible and capture multimodality in the latent space. Moreover from Eq. (4), our *mAR* prior can model long-range correlations in the latent space as the distribution of each channel is dependent on all previous channels.

This autoregressive factorization allows us to employ Conv-LSTMs (Shi et al., 2017) to model the distributions $p_\phi(\mathbf{l}_i^j|\mathbf{l}_i^1, \cdots, \mathbf{l}_i^{j-1}, \mathbf{r}_i)$ and $p_\phi(\mathbf{h}_n)$. Conv-LSTMs can model long-range dependencies across channels in their internal state. Additionally, long-range spatial dependencies within channels can be modeled by stacking multiple Conv-LSTM layers with a wide receptive field. This formulation allows all pixels within a channel to be sampled in parallel, while the channels are sampled in a sequential manner,

$$\hat{\mathbf{l}}_i^j \sim p_\phi\Big(\mathbf{l}_i^j \Big| \mathbf{l}_i^1, \cdots, \mathbf{l}_i^{j-1}, \mathbf{r}_i\Big). \quad (6)$$

This is in contrast to PixelCNN/RNN-based models, which sample one pixel at a time.

The *mAR-SCF* model. We illustrate our *mAR-SCF* model architecture in Fig. 1. Our *mAR-SCF* model leverages the SQUEEZE and SPLIT operations for invertible flows introduced in (Dinh et al., 2015; 2017) for efficient parallelization. Following (Dinh et al., 2015; 2017; Kingma & Dhariwal, 2018), we use several SQUEEZE and SPLIT operations in a multi-scale setup at n scales until the spatial resolution at $\mathbf{h}_n$ is reasonably small, typically 4×4. Note that there is no SPLIT operation at the smallest spatial resolution. Therefore, the latent space is the concatenation of $\mathbf{z} = \{\mathbf{l}_1, \ldots, \mathbf{l}_{n-1}, \mathbf{h}_n\}$. The split coupling flows (SCF) f_{θ_i} in the *mAR-SCF* model remain invertible by construction. We consider different SCF couplings for f_{θ_i}, including the affine couplings of (Dinh et al., 2017; Kingma & Dhariwal, 2018) and MixLogCDF couplings (Ho et al., 2019).

Given the parameters ϕ of our multimodal *mAR* prior modeled by the Conv-LSTMs, we can compute $p_\phi(\mathbf{z})$ using the formulation in Eqs. (4) and (5). We can thus express Eq. (3) in *closed form* and directly maximize the likelihood of the data under the multimodal *mAR* prior distribution.

Analysis of sampling time. We now formally analyze the computational cost in the number of steps T required for sampling with our *mAR-SCF* model. First, we describe the sampling process in detail in Algorithm 1 (the forward training process follows the sampling process in reverse order). Next, we derive the worst-case number of steps T required by MARPS, given sufficient parallel resources to sample a channel in parallel. Here, the number of steps T can be seen as the length of the critical path while sampling.

Algorithm 1 MARPS: Multi-scale Autoregressive Prior Sampling for our *mAR-SCF* models

Sample $\hat{\mathbf{h}}_n \sim p_\phi(\mathbf{h}_n)$
for $i \leftarrow n-1, \cdots, 1$ **do**
  /* SPLITINVERSE */
  $\hat{\mathbf{r}}_i \leftarrow \hat{\mathbf{h}}_{i+1}$; // Assign previous
  $\hat{\mathbf{l}}_i \sim p_\phi(\mathbf{l}_i|\mathbf{r}_i)$; // Sample *mAR* prior
  $\hat{\mathbf{h}}_i \leftarrow \Big\{\hat{\mathbf{l}}_i, \hat{\mathbf{r}}_i\Big\}$; // Concatenate
  /* STEPOFFLOWINVERSE */
  Apply $f_i^{-1}(\hat{\mathbf{h}}_i)$; // SCF coupling
  /* SQUEEZEINVERSE */
  Reshape $\hat{\mathbf{h}}_i$; // Depth to Space
end for
$\mathbf{x} \leftarrow \hat{\mathbf{h}}_1$

Lemma 3.1. *Let the sampled image* $\mathbf{x}$ *be of resolution* $[C, N, N]$*, then the worst-case number of steps* T *(length of the critical path) required by MARPS is* $\mathcal{O}(N)$.

We include the proof in the Appendix. It follows that with our multi-scale autoregressive *mAR* priors in our *mAR-SCF* model, sampling can be performed in a linear number of time-steps in contrast to fully autoregressive models like PixelCNN, which require a quadratic number of time-steps (van den Oord et al., 2016b).

4. Experiments

We evaluate our approach on the CIFAR10 (Krizhevsky et al., 2009), and on ImageNet (van den Oord et al., 2016b) in the Appendix. We provide detailed architecture and hyperparameter details in the Appendix [*].

Density estimation. The density estimation results in terms of bits/dim using the per-pixel log-likelihood metric on CIFAR10 are shown in Table 1. We include the architecture details (# of levels, coupling type, # of channels). We also report the sampling speed in terms of sampling one image using an average over 1000 runs and a batch size of 32. We compare to the state-of-the-art exact inference methods – Glow (Kingma & Dhariwal, 2018) and Flow++ (Ho et al., 2019) methods with SCF couplings and Residual Flows (Chen et al., 2019). For fair comparison with Glow and Residual flows, we use uniform dequantization unlike Flow++, which proposes to use variational dequantization.

In comparison to Glow, we achieve improved density estimation results. In detail, we outperform Glow (3.35 *vs*. 3.33

[*]Code: https://github.com/visinf/mar-scf

Method	Coupling	Levels	|SCF|	Channels	**bits/dim** (↓)	**Speed** (ms, ↓)
PixelCNN (van den Oord et al., 2016b)	Autoregressive	–	–	–	3.00	4×10^3
PixelCNN++ (van den Oord et al., 2016a)	Autoregressive	–	–	–	*2.92*	5×10^3
Glow (Kingma & Dhariwal, 2018)	Affine	3	32	512	3.35	13
Flow++ (Ho et al., 2019)	MixLogCDF	3	–	96	3.29	19
Residual Flow (Chen et al., 2019)	Residual	3	16	–	3.28	34
mAR-SCF (Ours)	Affine	3	32	256	3.33	**6**
mAR-SCF (Ours)	Affine	3	32	512	3.31	17
mAR-SCF (Ours)	MixLogCDF	3	4	96	3.27	19
mAR-SCF (Ours)	MixLogCDF	3	4	256	**3.22**	32

Table 1. Evaluation of our *mAR-SCF* model on CIFAR10 (using uniform dequantization for fair comparsion with (Chen et al., 2019; Kingma & Dhariwal, 2018)).

	PixelCNN[1]	PixelIQN[1]	Glow[2]	Residual Flow[2]	*mAR-SCF* (Ours)	ANF[3]	DCGAN[4]	WGAN-GP[4]
FID (↓)	65.9	49.4	46.9	46.3	33.6	30.6	37.1	*29.3*
IS (↑)	4.6	–	–	5.2	6.4	6.5	6.4	*7.8*

Table 2. Evaluation of sample quality on CIFAR10. Other results are quoted from (Ostrovski et al., 2018)[1], (Chen et al., 2019)[2], (Huang et al., 2020)[3] (Gulrajani et al., 2017; Heusel et al., 2017)[4]

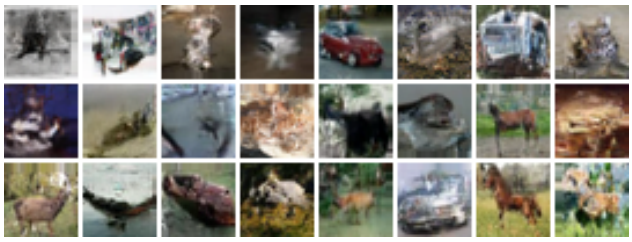


Figure 2. Random samples, *mAR-SCF* with MixLogCDF couplings (3.22 bits/dim, 33.6 FID).

bits/dim) with |SCF|= 32 affine couplings and 3 levels, while using parameter prediction networks with only half (256 *vs*. 512) the number of channels. We observe that increasing the capacity of our parameter prediction networks to 512 channels boosts the log-likelihood further to 3.31 bits/dim. As this setting with 512 channels is identical to setting reported in (Kingma & Dhariwal, 2018), it serves as an ideal ablation. We see that our *mAR* prior boosts the accuracy by ~ 0.04 bits/dim. This performance gain is competitive with the ~ 0.03 bits/dim boost reported by Glow (*cf*. Fig. 3 in (Kingma & Dhariwal, 2018)) with the introduction of the 1×1 convolution. We train our model for ~ 3000 epochs, similar to Glow. Also note that we only require a batch size of 128 to achieve state-of-the-art likelihoods, whereas Glow uses batches of size 512. Thus our *mAR-SCF* model improves density estimates and requires significantly lower computational resources (~ 48 *vs*. ~ 128 GB memory). Overall, we also observe competitive sampling speed. This firmly establishes the utility of our *mAR-SCF* model.

For fair comparison with Flow++ (Ho et al., 2019) and Residual Flows (Chen et al., 2019), we employ the more powerful MixLogCDF couplings. Our *mAR-SCF* model uses 4 MixLogCDF couplings at each level with 96 channels similar to Flow++ but includes SPLIT operations unlike Flow++. Here, we outperform Flow++ and Residual Flows (3.27 *vs*. 3.29 and 3.28 bits/dim) while being equally fast to sample as Flow++. A baseline model without our *mAR* prior has performance comparable to Flow++ (3.29 bits/dim). Finally, we train a more powerful *mAR-SCF* model with 256 channels with sampling speed competitive with (Chen et al., 2019), which achieves state-of-the-art 3.24 bits/dim on CIFAR10 after ~ 400 training epochs (comparable to ~ 350 epochs required by (Chen et al., 2019) to achieve 3.28 bits/dim). Training our model for $\sim$1000 epochs until convergence further improves results to 3.22 bits/dim.

Sample quality. We analyze the sample quality of our *mAR-SCF* model in Table 2 using the FID metric (Heusel et al., 2017) and Inception scores (Salimans et al., 2016). We achieve an FID of 33.6 and an Inception score of 6.4 with our *mAR-SCF* model with MixLogCDF couplings trained for $\sim$1000 epochs. We obtain improved sample quality over exact inference approaches and close the gap in comparison to augmented flows (Huang et al., 2020) and adversarial approaches like DCGAN and WGAN-GP. We show samples in Fig. 2 and in the Appendix.

5. Conclusion

We presented *mAR-SCF*, a flow-based generative model with novel multi-scale autoregressive priors for modeling long-range dependencies in the latent space of flow models. Our *mAR-SCF* model not only improves density estimation, but also considerably improves the sample quality compared to

previous state-of-the-art exact inference models. We believe the combination of complex priors with flow-based models, as demonstrated by our *mAR-SCF* model, provides a path toward efficient models for exact inference that approach the fidelity of GAN-based approaches.

Acknowledgement. SM and SR acknowledge the support by the German Research Foundation as part of the Research Training Group *Adaptive Preparation of Information from Heterogeneous Sources (AIPHES)* under grant No. GRK 1994/1.

References

Chen, R. T. Q., Behrmann, J., Duvenaud, D., and Jacobsen, J. Residual flows for invertible generative modeling. In *NeurIPS*, pp. 9913–9923, 2019.

Dinh, L., Krueger, D., and Bengio, Y. NICE: Non-linear independent components estimation. In *ICLR Workshop*, 2015.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. In *ICLR*, 2017.

Domke, J., Karapurkar, A., and Aloimonos, Y. Who killed the directed model? In *CVPR*, 2008.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *NeurIPS*, pp. 5767–5777, 2017.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, pp. 6626–6637, 2017.

Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *ICML*, pp. 2722–2730, 2019.

Huang, C., Dinh, L., and Courville, A. C. Augmented normalizing flows: Bridging the gap between generative flows and latent variable models. *CoRR*, 2020. URL https://arxiv.org/abs/2002.07101.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, pp. 10215–10224, 2018.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, U. of Toronto, 2009.

Ostrovski, G., Dabney, W., and Munos, R. Autoregressive quantile networks for generative modeling. In *ICML*, pp. 3936–3945, 2018.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

Razavi, A., van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, pp. 14837–14847, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(13):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. In *NeurIPS*, pp. 2234–2242, 2016.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D., Wong, W., and Woo, W. Deep learning for precipitation nowcasting: A benchmark and a new model. In *NeurIPS*, pp. 5617–5627, 2017.

van den Oord, A., Kalchbrenner, N., Espeholt, L., Kavukcuoglu, K., Vinyals, O., and Graves, A. Conditional image generation with PixelCNN decoders. In *NeurIPS*, pp. 4790–4798, 2016a.

van den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. In *ICML*, pp. 1747–1756, 2016b.

Wei, X., Gong, B., Liu, Z., Lu, W., and Wang, L. Improving the improved training of Wasserstein GANs: A consistency term and its dual effect. In *ICLR*, 2018.

Ziegler, Z. M. and Rush, A. M. Latent normalizing flows for discrete sequences. In *ICML*, pp. 7673–7682, 2019.

A. Appendix

We provide additional details of Lemma 3.1 in the main paper, additional architecture and hyperparameters details of our *mAR-SCF* model, as well as additional results and qualitative examples.

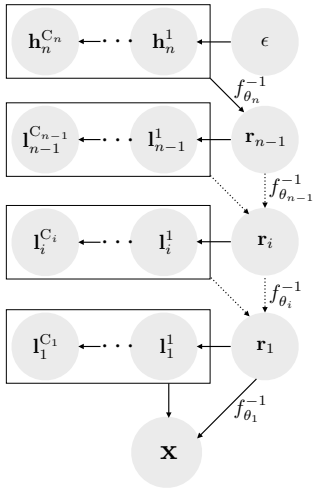


Figure 3. Generative model for mAR-SCF. The generative model shows the multi-scale autoregressive sampling of the channel dimensions of $\mathbf{l}_i$ at each level. The spatial dimensions of each channel are sampled in parallel. $\mathbf{r}_i$ are computed with invertable transformations.

A.1. Additional Details of Lemma 3.1

We derive the worst-case number of steps T required by MARPS, given sufficient parallel resources to sample a channel in parallel. Here, the number of steps T can be seen as the length of the critical path while sampling.

Lemma A.1. *Let the sampled image* $\mathbf{x}$ *be of resolution* $[C, N, N]$*, then the worst-case number of steps* T *(length of the critical path) required by MARPS is* $\mathcal{O}(N)$.

Proof. At the first sampling step (Fig. 3) at layer f_{θ_n}, our *mAR* prior is applied to generate $\mathbf{h}_n$, which is of shape $[2^{n+1}\,C, N/2^n, N/2^n]$. Therefore, the number of sequential steps required at *the last flow layer* $\mathbf{h}_n$ is

$$T_n = C \cdot 2^{n+1}. \tag{7}$$

Here, we are assuming that each channel can be sampled in parallel in one time-step.

From $f_{\theta_{n-1}}$ to f_{θ_1}, f_{θ_i} always contains a SPLIT operation. Therefore, at each f_{θ_i} we use our *mAR* prior to sample $\mathbf{l}_i$, which has shape $[2^i\,C, N/2^i, N/2^i]$. Therefore, the number of sequential steps required for sampling at layers $\mathbf{h}_i, 1 \leq i < n$ of our *mAR-SCF* model is

$$T_i = C \cdot 2^i. \tag{8}$$

Therefore, the total number of sequential steps (length of the critical path) required for sampling is

$$\begin{aligned} T &= T_n + T_{n-1} + \cdots + T_i + \cdots + T_1 \\ &= C \cdot \left(2^{n+1} + 2^{n-1} + \cdots + 2^i + \cdots + 2^1\right) \\ &= C \cdot \left(3 \cdot 2^n - 2\right). \end{aligned} \tag{9}$$

Now, the total number of layers in our *mAR-SCF* model is $n \leq \log(N)$. This is because each layer reduces the spatial resolution by a factor of two. Therefore, the total number of time-steps required is

$$T \leq 3 \cdot C \cdot N. \tag{10}$$

In practice, $C \ll N$, with $C = \mathrm{C}_0 = 3$ for RGB images. Therefore, the total number of sequential steps required for sampling in our *mAR-SCF* model is $T = \mathcal{O}(N)$. □

A.2. Implementation Details

Our *mAR* prior at each level f_{θ_i} consists of three convolutional LSTM layers, each of which uses 32 convolutional filters to compute the input-to-state and state-to-state components. Keeping the *mAR* prior architecture constant, we experiment with different SCF couplings in f_{θ_i} to highlight the effectiveness of our *mAR* prior. We experiment with affine couplings of (Dinh et al., 2017; Kingma & Dhariwal, 2018) and MixLogCDF couplings (Ho et al., 2019). Affine couplings have limited modeling flexibility. The more expressive MixLogCDF applies the cumulative distribution function of a mixture of logistics. In the following, we include experiments varying the number couplings and the number of channels in the convolutional blocks of the neural networks used to predict the affine/MixLogCDF transformation parameters.

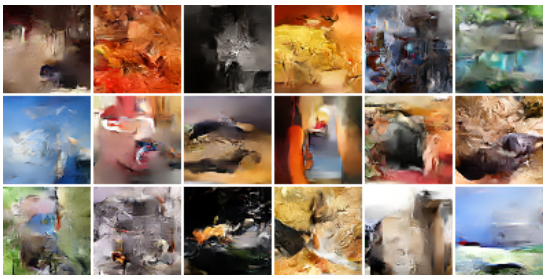
We use Adamax (as in (Kingma & Dhariwal, 2018)) with a learning rate of 8×10^{-4}. We use a batch size of 128 with affine and 64 with MixLogCDF couplings (following (Ho et al., 2019)).

A.3. ImageNet

We evaluate our *mAR-SCF* model on ImageNet (32×32 and 64×64) against the best performing models on CIFAR10 in Table 3, i.e. Glow (Kingma & Dhariwal, 2018) and Residual Flows (Chen et al., 2019). Our model with affine couplings outperforms Glow while using fewer channels (4.07 *vs*. 4.09 bits/dim). For comparison with the more powerful Residual Flow models, we use four MixLogCDF couplings at each layer f_{θ_i} with 460 channels. We again outperform Residual Flows (Chen et al., 2019) (3.99 *vs*. 4.01 bits/dim). These results are consistent with the findings in Table 1, highlighting the advantage of our *mAR* prior. Finally, we also evaluate on the ImageNet (64×64) dataset. Our *mAR-SCF* model with

Method	Coupling	|SCF|	Ch.	**bits/dim** (↓)	**Mem (GB,** ↓)
Glow (Kingma & Dhariwal, 2018)	Affine	32	512	4.09	128
Residual Flow (Chen et al., 2019)	Residual	32	–	4.01	–
mAR-SCF (Ours)	Affine	32	256	4.07	**48**
mAR-SCF (Ours)	MixLogCDF	4	460	**3.99**	80

Table 3. Evaluation on ImageNet (32×32).



(a) Residual Flows (Chen et al., 2019) (3.75 bits/dim)

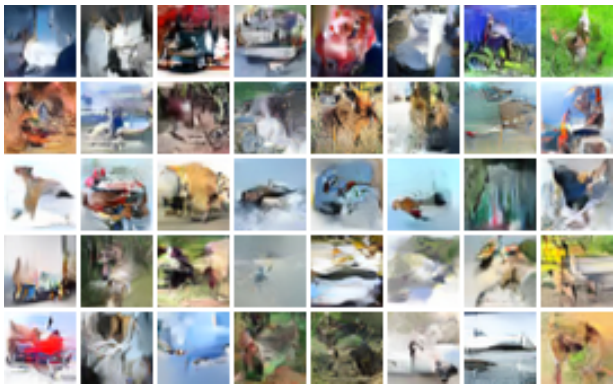
(b) Our *mMAR-SCF* (Affine, 3.80 bits/dim)

Figure 4. Random samples on ImageNet (64×64).

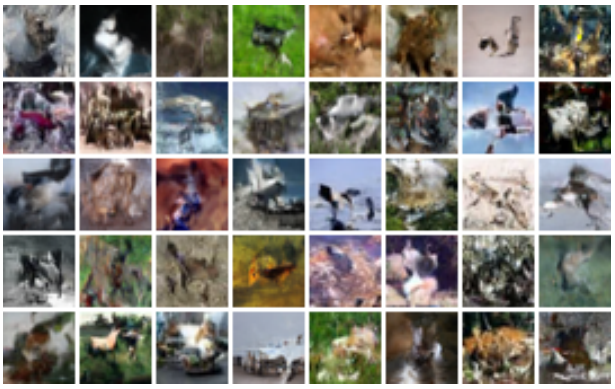
affine flows achieves 3.80 *vs*. 3.81 bits/dim in comparison to Glow (Kingma & Dhariwal, 2018). We show qualitative examples in Fig. 4 and compare to Residual Flows. We see that although the powerful Residual Flows obtain better log-likelihoods (3.75 bits/dim), our *mAR-SCF* model achieves better visual fidelity. This again highlights that our *mAR* is able to better capture long-range correlations.

A.4. Additional Qualitative Examples

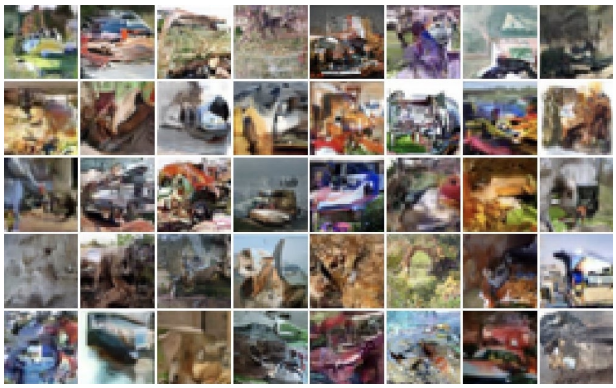
We include additional qualitative examples on CIFAR10 in Fig. 5. We compare with the fully autoregressive PixelCNN model (van den Oord et al., 2016b), Flow++ (Ho et al., 2019) and and Residual flows (Chen et al., 2019). We see that our *mAR-SCF* models achieves better visual sample quality (also shown by the FID and Inception metrics in Fig. 2 of the main paper), with more clearly defined objects.



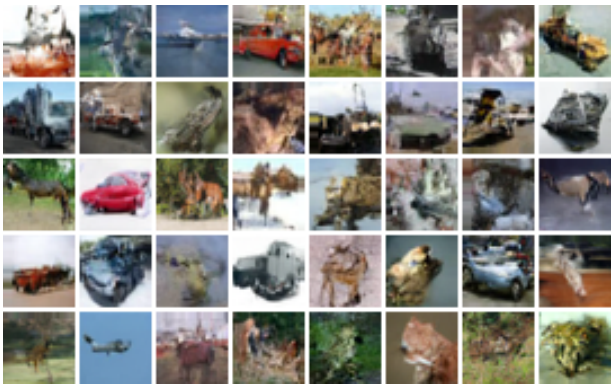
(a) Residual Flows (Chen et al., 2019) (3.28 bits/dim, 46.3 FID)



(b) Flow++ with variational dequantization (Ho et al., 2019) (3.08 bits/dim)



(c) PixelCNN (van den Oord et al., 2016b) (3.00 bits/dim)



(d) Our *mMAR-SCF* MixLogCDF (3.22 bits/dim)

Figure 5. Comparison of random samples from our *mAR-SCF* model with state-of-the-art models on CIFAR10.