Table 1: (a) Image synthesis speed on CIFAR10. Glow re-implemented in PyTorch is masked with †. ‡ denotes results shown in Hoogeboom et al. [2019]. (b) Image synthesis speed of MaCow on different datasets.

(a)

CIFAR10	time/sample (ms)	Slow-down
Glow‡	5	1.0
MAF ‡	3000	600.0
Emerging‡	1800	360.0
Glow†	5.3	1.0
MaCow	38.7	7.3

(b)

Dataset	image size	time/sample (ms)
CIFAR10	32×32	38.7
ImageNet-64	64×64	104.7
LSUN-128	128×128	267.9
CelebA-HQ-256	256×256	434.2

We thank for the valuable feedback. We address the questions below and will revise our paper accordingly.

[R1 & R2 & R3 Generation speed] We appreciate your suggestions on adding experiments on image generation speed comparison and plan to add a section in experiments to discuss it in details. Table 1a and Table 1b provides the preliminary results. Following Hoogeboom et al. [2019], we measure the time to sample a datapoint when computed in mini-batchs with size 100. For fair comparison, we re-implemented Glow using PyTorch and show that the speed of Glow† and Glow‡ is comparable. On CIFAR-10, MaCow is 7.3 times slower than Glow, much faster than Emerging Convolution and MAF, whose factors are 360 and 600 respectively. The generation speed of MaCow on different datasets is shown in Table 1b. We see that the time of generation increases linearly with the the image resolution.

[R2 Related work and Originality] Thanks for pointing out the related work. We will cite the Emerging Convolution paper in our final version and discuss the detailed relation and difference with MaCow. There are two main differences between MaCow and Emerging Convolution. i) the pattern of the mask is different. With the mask in MaCow (Figure 1 in the paper), MaCow achieves significantly more efficient inference and sampling (shown in Table 1a and 1b) by reducing the complexity from $O(h \times w)$ to $O(h)$ or $O(w)$, without sacrificing the receptive field. ii) the Emerging Convolutional Flow [Hoogeboom et al., 2019] is basically a linear transformation with masked convolutional kernels, which does not introduce "nonlinearity" to the random variables. This flow is proposed as an alternative to the 1×1 conv flow in Glow. MaCow, in contrast, is able to introduce such nonlinearity similar to the coupling layer, and is proposed to replace both the coupling layers and the 1×1 conv layers in Glow. We would like to point out that your comment "the masked conv layers in MaCow are linear" is a misconception. Actually, regardless of the different patterns of masks, Emerging Conv is a special (linear) case of MaCow, by specifying the $s(x_{<t})$ and $b(x_{<t})$ in Eq (4) as linear functions.

[R2 & R3 Novelty and significance of improvements] First, we would like to argue that the proposed new pattern of mask in MaCow is simple and effective, but not trivial or incremental. It produces a semi-autoregressive model, which significantly reduces the inference time of MAF and obtains better density estimation performance.

Second, the improvements of MaCow on bits/dim, especially the contribution of fine-grained multi-scale architecture, are not neglectable. Emerging Convolution [Hoogeboom et al., 2019] obtained 0.02 improvement on bits/dim by increasing both the number of parameters and inference time. Our fine-grained architecture reduces the number of parameters (compared with the original multi-scale architecture) and obtains 0.03 bit/dim improvements on both CIFAR-10 and ImageNet-64. The overall improvements of MaCow over Glow, without Variational dequantization, are 0.07 on CIFAR-10, 0.06 on ImageNet-64, 0.04 on LSUN-bedroom and 0.08 on CelebA-HQ.

[R3 Flow++ v.s. MaCow] From results shown in Table 1 of our submission, in terms of density estimation, we could find that on CIFAR-10 with uniform dequantization, MaCow (3.28) performs better than Flow++ (3.29) and on ImageNet-64, with variational dequantization, MaCow (3.66) outperforms Flow++ (3.69). The only exception is on CIFAR-10 with variational dequantization, Flow++ (3.09) achieve better performance than Macow (3.16). But we have to mention that even with similar number of parameters, Flow++ is slower and consumes much more memory than Glow and MaCow, preventing us from evaluating it on high-resolution images.

[R3 Sampling with temperature] The temperature trick is only applied to LSUN and CelebA-HQ 5-bits images, where MaCow adopts additive coupling layers. For CIFAR-10 and ImageNet 8-bits images, we sample with temperature 1.0.

[R1 non-numerical measures of modifications] We appreciate your suggestion on evaluating the difference between MaCow w./w.o. these modications in non-numerical ways. From the image samples w./w.o the variational dequantization, we have not observed significant difference. We will consider some other non-numerical metrics.

[R2 & R3 Error bar] We have performed experiments on CIFAR-10 with multiple runs (more than 3) with different random seeds and the standard deviation is less than 0.005. On other datasets, same with Glow and Flow++, we only performed single run on each dataset due to the limits of computational resources, unfortunately.

[R2 statement of stability] We appreciate your comment about our irrigorous statement of stability in line 128. We will amend this claim in the final version.

[R1 & R2 & R3 Paper writing] We thank for all your suggestions on revising the paper to improve the writing. We will elaborate the Masked Convolution and Fine-grained architecture sections to explain them more clearly.

References

Emiel Hoogeboom, Rianne V. Berg, and Max Welling. Emerging convolutions for generative normalizing flows. In *ICML*, 2019.