

Memory Efficient Diffusion Probabilistic Models via Patch-based Generation

Shinei Arakawa¹ Hideki Tsunashima¹ Daichi Horita² Keitaro Tanaka¹ Shigeo Morishima³
¹Waseda University ²The University of Tokyo
³Waseda Research Institute for Science and Engineering

Abstract

Diffusion probabilistic models have been successful in generating high-quality and diverse images. However, traditional models, whose input and output are high-resolution images, suffer from excessive memory requirements, making them less practical for edge devices. Previous approaches for generative adversarial networks proposed a patch-based method that uses positional encoding and global content information. Nevertheless, designing a patch-based approach for diffusion probabilistic models is non-trivial. In this paper, we present a diffusion probabilistic model that generates images on a patch-by-patch basis. We propose two conditioning methods for a patch-based generation. First, we propose position-wise conditioning using one-hot representation to ensure patches are in proper positions. Second, we propose Global Content Conditioning (GCC) to ensure patches have coherent content when concatenated together. We evaluate our model qualitatively and quantitatively on CelebA and LSUN bedroom datasets and demonstrate a moderate trade-off between maximum memory consumption and generated image quality. Specifically, when an entire image is divided into 2×2 patches, our proposed approach can reduce the maximum memory consumption by half while maintaining comparable image quality.

1. Introduction

Diffusion Probabilistic Models (DPMs) [5, 22, 23] have achieved remarkable success in high-resolution image synthesis [2, 15], image editing [1, 14], and super-resolution [20]. Stable Diffusion (SD) [18] is a popular text-to-image diffusion model trained on a large-scale text-image paired dataset, LAION5B [21]. SD outperforms other generative models, including GANs [3, 8, 17] and VAEs [10], in both generative quality and diversity. Despite being able to generate images on commercially available GPUs, SD faces challenges when generating high-resolution images of 512^2 or executing on edge devices such as smartphones due to GPU memory limitations.

To address the similar limitation for GANs, COCO-

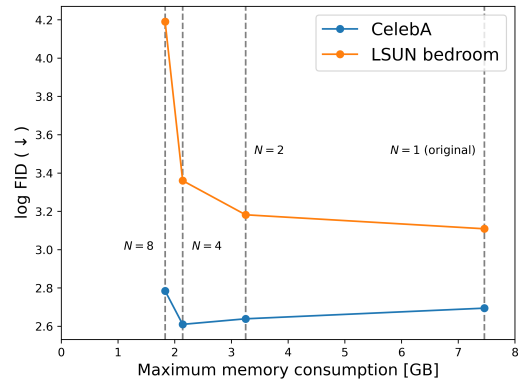


Figure 1. Trade-off between the maximum memory consumption and the generative quality measured by (log) FID. N indicates the number of patch divisions for both vertical and horizontal directions. Our proposed method demonstrates moderate trade-off.

GAN [12] proposed a memory-efficient approach by parallelly generating patches and concatenating them to generate a whole image. COCO-GAN introduced a position embedding mechanism to specify patch locations and an information-sharing mechanism to ensure coherent image content across patches. However, memory-efficient methods for DPMs have not been sufficiently discussed.

In this study, we present a DPM that generates images on a patch-by-patch basis to reduce memory consumption during inference. Patch-based generation effectively reduces the size of the input to the self-attention mechanism [24], which is commonly used in the network of DPMs and can be a bottleneck for memory consumption. Consequently, patch-based generation is capable of significantly reducing the maximum amount of memory required for computation.

Specifically, our study introduces two novel conditioning methods to enable DPMs to generate images patch-by-patch. The first method conditions location information that specifies the position of the generated patch within the entire image. We embed the position using a one-hot representation to enable proper conditioning. The second method conditions the content information of the entire image to ensure coherence in the generated patches when rearranged. To embed content information, we propose Global Content Conditioning (GCC), which first downsamples the entire

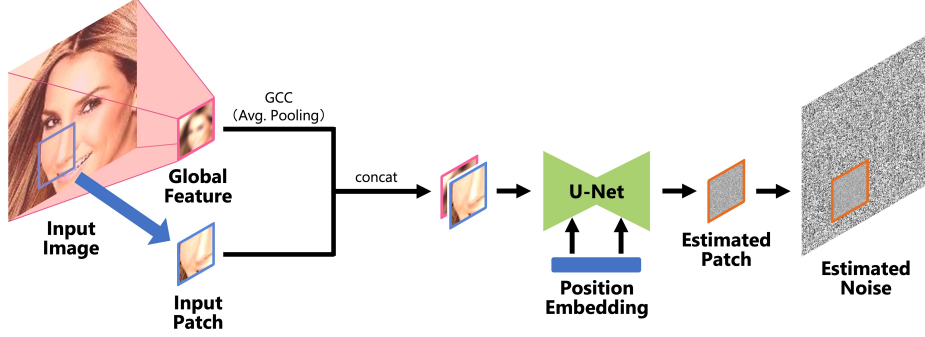


Figure 2. An overview of our proposed method. Global Content Conditioning (GCC) extracts the global content feature from the entire image. After that, the global content feature is concatenated with the cropped patch and fed into the U-Net conditioned by patch position embedding. Our approach enables DPMs to generate images patch-by-patch, achieving high memory efficiency during inference.

image to get content information and then conditions the network by concatenating the extracted content feature with patches. GCC can be implemented without any additional trainable parameters. Through these conditioning methods, the proposed approach significantly reduces memory consumption during inference without dramatically increasing the number of network parameters. An overview of our proposed method is illustrated in Fig. 2.

We evaluated our proposed method on CelebA [13] and LSUN bedroom [25] datasets, which consist of face and bedroom images, respectively. Our experimental results shed light on the trade-off relationship between generation quality and maximum memory consumption during inference, shown in Fig. 1. Specifically, our approach can reduce the maximum memory consumption by half while maintaining comparable image quality when dividing an entire image into 2×2 patches.

2. Background

Sohl-Dickstein et al. [22] introduced Diffusion Probabilistic Models (DPMs) as a deep generative method. Ho et al. [5] subsequently extended the model and achieved high generative performance on diverse datasets (e.g., CIFAR10 [11], CelebA-HQ [6], and LSUN [25]). DPMs consist of forward and reverse processes. In this section, we provide a detailed description of each process.

Forward Process In DPMs, an image \mathbf{x}_0 sampled from data distribution $p(\mathbf{x})$ is perturbed iteratively by adding Gaussian noise. We define each diffusion step as $t \in \{1, \dots, T\}$, with the perturbed latent image denoted as \mathbf{x}_t . The gaussian transition is given by,

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where β_t is the variance parameter at step t , determined such that $\mathbf{x}_T \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$. Using Eq. (1), we can write the conditional distribution of the latent variable \mathbf{x}_t , given the

raw image \mathbf{x}_0 , as a closed form,

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (2)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. Given a real image \mathbf{x}_0 , we can compute the perturbed image \mathbf{x}_t at any diffusion step. This transition from data distribution to Gaussian distribution through repeated Gaussian noise addition constitutes the forward process.

Reverse Process The reverse process reproduces the data distribution by iterative denoising from a Gaussian distribution. Starting from $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, the iterative denoising is performed using the equation,

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}, \quad (3)$$

where $\sigma_t = \sqrt{\beta_t}$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The noise added to the latent image \mathbf{x}_t is approximated by a neural network denoted as $\epsilon_\theta(\mathbf{x}_t, t)$. U-Net [19] is commonly used for the neural network $\epsilon_\theta(\mathbf{x}_t, t)$.

Loss Function The loss function for DPMs is expressed as a squared error,

$$L = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} \left[\left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t) \right\|^2 \right]. \quad (4)$$

In this formulation, the network of DPMs is trained by minimizing the squared error between the added noise in the forward process and the estimated noise.

3. Proposed Method

The network used in DPMs requires a significant amount of memory for computation since both input and output are high-dimensional images. Our objective in this study is to reduce memory consumption during inference by partitioning images into patches and reducing the size of input and output variables passed to the network. In this section, we describe a two-step approach for patch-by-patch generation.

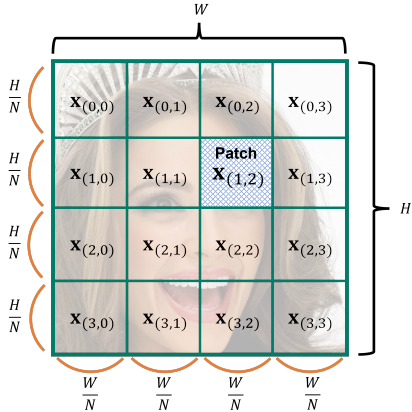


Figure 3. An example of the patch partitioning ($N = 4$). The entire image is divided into $N^2 = 16$ patches, without any overlap.

First, we introduce a method for extracting patches from the entire image. Next, we propose two conditioning methods that provide each location and content information required for generating patches.

3.1. Patch Partitioning Method

In conventional DPMs, the entire noisy image is passed into the network, and the amount of noise is estimated by a single forward propagation. Our patch-by-patch generation method divides the input whole image into patches, passes the patches to the network individually, and rearranges the output patches to obtain the overall output. Specifically, the input image \mathbf{x} is divided into patches $\mathbf{x}_{(i,j)}$ in a grid-like manner with no overlap. An example of the proposed patch partitioning method is shown in Fig. 3. We define the resolution of the input image \mathbf{x} as (H, W) . If the image is divided into N equal parts in each direction, the resolution of a patch is expressed as $(H', W') = (H/N, W/N)$. With the number of divisions N , the range of indices i and j is defined as $0 \leq i, j \leq N - 1$.

3.2. Conditioning Location Information

As the divided patches are independently input into the network, it is essential to provide patch location information to the network. We address this by defining an index s to specify the position of each patch uniquely. We define the index s for N -divided patches $\mathbf{x}_{(i,j)}$ as $s = i \times N + j$. After converting the index s to a one-hot representation (e.g., $[0, 0, 1, 0]$), it is passed to a fully connected layer, and the output is concatenated with the embedding features of the diffusion step (corresponding to the position embedding in Fig. 2) and fed into the network. This approach conditions the network with the relative position of an input patch.

3.3. Conditioning Content Information

We further propose Global Content Conditioning (GCC) to address the lack of content information shared between

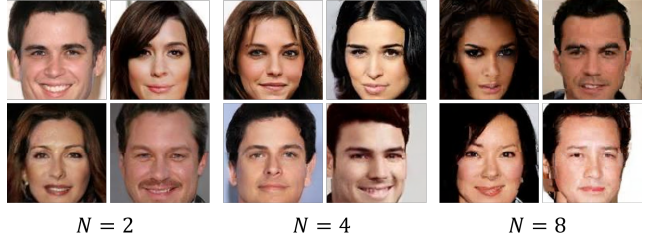


Figure 4. Generated images from our proposed method trained on CelebA dataset [13].



Figure 5. Generated images from our proposed method trained on LSUN bedroom dataset [25].

independently processed patches in the network. GCC extracts the overall content information $\mathbf{g} \in \mathbb{R}^{H' \times W'}$ through average pooling of the entire image $\mathbf{x} \in \mathbb{R}^{H \times W}$. Then \mathbf{g} is concatenated with the split patch $\mathbf{x}_{(i,j)}$ in the channel direction and input into the network. GCC is parameter-free, and it can be implemented without adding model complexity. This method is motivated by the technique proposed by Saharia et al. [20], which demonstrates effective conditioning through simple concatenation in the channel direction for conditioning image features.

4. Experiments

4.1. Experiment Settings

Dataset We conducted experiments on two datasets: CelebA [13], consisting of approximately 200K face images, and LSUN bedroom [25], consisting of approximately three million bedroom images. We preprocessed all images to be 128×128 pixels in size. For CelebA, as the images were already square, we just downsampled them to 128^2 . For LSUN bedroom, we first cropped the images with the short side resolution and then downsampled the cropped images to 128^2 . The datasets were split into train, validation, and test sets, with proportions of 64% : 16% : 20%.

Training The number of diffusion steps T was fixed at 1000. The mini-batch sizes were 64 for both CelebA and LSUN bedroom. The learning rate was set to 1×10^{-4} , and we chose Adam [9] as an optimization method. We updated the parameters for 800,000 iterations on both datasets. We experimented with different numbers of patch divisions N , specifically $N = 2, 4, 8$, and also trained the conventional DPM without patch partitioning as the original method.

Table 1. FID (\downarrow) evaluation of our proposed methods and the original. The lower is better.

Dataset	Division N	FID (\downarrow)
CelebA	1 (original)	14.8
	2	14.0
	4	13.6
	8	16.2
LSUN bedroom	1 (original)	22.4
	2	24.1
	4	28.8
	8	66.1

Table 2. Maximum memory consumption of the original method and our proposed method at different number of divisions N .

Division N	Max memory consumption [GB]
1 (original)	7.46
2	3.25
4	2.14
8	1.83

Evaluation Metrics We used Fréchet Inception Distance (FID) [4, 16], a metric that measures the distance between two image distributions and provides a measure of generative quality and diversity. For CelebA, we used 40K generated images and 40K real images to measure the FID score. For LSUN bedroom, we used 50K generated images and 50K real images. Real images were selected randomly from each test set. Additionally, we used the maximum memory consumption during inference as another evaluation metric.

4.2. Qualitative Evaluation

The results of the proposed method trained on CelebA dataset are presented in Fig. 4. When $N = 2$ or 4, the generated images appear natural. However, when $N = 8$, the boundary lines between patches become prominent, even though content information is shared among the patches. Similarly, Fig. 5 shows the results of the proposed method on LSUN bedroom dataset. The generated images look natural when $N = 2$, but when $N \geq 4$, patch boundaries become noticeable. While the network is conditioned with location information, the conditioning is not precise enough to achieve seamless patch connection at the boundaries, where accurate location information is essential.

4.3. Quantitative Evaluation

The proposed method’s FID results are presented in Tab. 1 for both CelebA and LSUN bedroom datasets. On the CelebA dataset, when $N = 2$ or 4, the proposed method’s performance is slightly better than the original method. However, as the number of patch divisions N increases up to eight, the FID score inclines sharply, consistent with the qualitative evaluation results. Similarly, for the LSUN bedroom dataset, the performance of the proposed method is

not significantly different from that of the original method when $N = 2$. However, the FID deteriorates sharply when the number of divisions increases over four or more.

The maximum memory consumption of the original and proposed methods is presented in Tab. 2. As the computation part of the reverse process is the same for both methods, only the forward propagation part of the network was considered for measuring the maximum memory consumption. The proposed method with $N = 2$ reduced the memory consumption by over 50% compared to the original method. When $N \geq 4$, the memory consumption decreases as N increases, although not uniformly due to the increased number of input channels caused by GCC. Fig. 1 shows the trade-off relation between maximum memory consumption and the FID score. Overall, the proposed method effectively reduces the maximum memory consumption during inference, preserving comparable generative quality.

5. Discussion

For LSUN bedroom dataset, significant degradation in quality was observed when the number of divisions N became four or more. In addition, comparing the results of CelebA and LSUN bedroom datasets for $N = 8$ divisions, LSUN bedroom showed a larger quality degradation compared to the original method. This is because the LSUN bedroom dataset includes various compositions and may be difficult to train compared to the CelebA dataset, which has a uniform composition. To overcome this issue, data augmentation techniques such as affine transformation proposed by Karras et al. [7] may be useful for datasets with non-uniform compositions.

Currently, GCC extracts content information from the entire input image at every diffusion step, which may lead to error accumulation during the reverse process and contribute to discontinuities on the boundaries between adjacent patches. To address this issue, further improvements can be made, such as limiting the extraction of the entire content information to only one time.

6. Conclusion

We introduced a patch-based DPM, a memory-efficient diffusion probabilistic model generating images on a patch-by-patch basis through positional embedding and Global Content Conditioning. The experiments showed that our approach could reduce the maximum memory consumption by half while maintaining comparable image quality when the entire image was divided into 2×2 patches.

Acknowledgements We thank Asst. Prof. Qi Feng for valuable feedbacks as well as English proofreading and Yoshiki Kubotani for feedbacks regarding the graphical design. This research is supported by the JSPS KAKENHI Grant Number 21H05054.

References

- [1] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *ICCV*, pages 14347–14356, 2021. 1
- [2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, volume 34, pages 8780–8794, 2021. 1
- [3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, volume 27, 2014. 1
- [4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, volume 30, 2017. 4
- [5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pages 6840–6851, 2020. 1, 2
- [6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 2
- [7] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 4
- [8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, June 2019. 1
- [9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3
- [10] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. 1
- [11] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 2
- [12] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. COCO-GAN: Generation by parts via conditional coordinating. In *ICCV*, pages 4511–4520, 2019. 1
- [13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 2, 3
- [14] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 1
- [15] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8162–8171, 2021. 1
- [16] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 4
- [17] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of stylegan imagery. In *ICCV*, pages 2085–2094, October 2021. 1
- [18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022. 1
- [19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 2
- [20] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023. 1, 3
- [21] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1
- [22] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2256–2265, 2015. 1, 2
- [23] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 1
- [24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017. 1
- [25] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015. 2, 3