Consistent Image Sequence Sampling from Single Image Diffusion Prior

Haojun Qiu

Abstract—Learning from a single instance, known as internal learning, has enabled diverse image processing tasks by leveraging patch statistics within a single image. Some deep generative methods, such as SinGAN [1], have demonstrated the ability to learn single-image patch distributions for tasks like super-resolution and image harmonization. However, these methods primarily focus on sampling individual images and do not address the challenge of generating consistent image sequences. In this work, we propose a novel training-sampling approach for single-image multi-scale diffusion models that enables consistent and joint sampling of image sequences. Our method bypass the traditional cascaded conditional super-resolution supervision by directly learning a patch-level prior across all scales of the training image parallelly in an image pyramid. At inference, we introduce a Laplacian-recomposition sampling algorithm to enforce multi-scale consistency and extend this framework to generate image sequences, such as infinite zoom-ins. We demonstrate the effectiveness of our approach in both single-image unconditional sampling and consistent image sequence generation.

1 Introduction

A recurring theme in machine learning is to learn from a large dataset of instances in an attempt to generalize. On the other hand, there has also been extensive research in **learning from a single instance**, a concept explored for over two decades. Referred to as internal learning, it uses only the patch statistics from a single instance, *e.g.*, an image, to accomplish various tasks. There are non-parametric regimes that boil down to searching for nearest neighbor patches within a single image using a distance metric like L2 distance. For example, non-local means methods [2] search for similar patches within an image and average them to denoise it, and super-resolution can be achieved by searching across image scales [3]. More generally, it has been shown that any natural image contains an extensive amount of patch recurrence [4], making the distribution of small patches from a single image a favorable subject of study. The concept of internal learning was later combined with advancements in deep learning. SinGAN [1] and InGAN [5] propose learning the patch distribution of a single image using a type of generative model—generative adversarial networks (GANs) [6]. In particular, SinGAN demonstrated the ability to learn a patch level prior from a single image, allowing for unconditional sampling from that image. Furthermore, these methods replicated success in tasks handled well by classical search-based methods, such as super-resolution [1] and image retargeting [5], *etc*..

- *M. Shell is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332.*
  E-mail: see http://www.michaelshell.org/contact.html
- *J. Doe is with Anonymous University.*

Fig. 1. Some fractal-like image from natural world.

All previous internal learning methods, however, have only studied sampling a single image (*e.g.*, a denoised, super-resolved, style-transferred, or harmonized image, etc.) with the single-image prior. A question remains unresolved is: **How can we sample a *sequence* of images, jointly and consistently, from a single-image prior?** To be more specific, an exciting task, for instance, is learning from a fractal-like image (some examples in fig. 1) that exhibits patch recurrence at all scales, with the goal of sampling a (pseudo-) infinite zoom-in sequence from it.

Previous single-image generative models could potentially be adapted for this new problem setting of image sequence sampling, but some problems exist. Following SinGAN [1], GPNN and SinDDM [7], [8], and several other works [9], [10], these methods are largely based on cascaded conditional super-resolution, where the generative model is supervised to conditionally and recursively scale up the sample from the coarsest to the finest scale of an image. Moreover, these methods do not inherently support generating a sequence of images in parallel. For tasks involving the sampling of a sequence of images, this type of conditional prediction within a single image (progressing over scales) as well as across different images falls short in ensuring consistency across images as well as preserving the patch statistics at all scales for each sampled image.

We propose a new training-sampling of diffusion models on a single image that enables consistent image sequence sampling. Specifically, our approach can be discussed in three stages. First, we learn the single image diffusion model. To sample an image such that its overlapping small patches are coming from the real distribution, our neural

network is of very small receptive field. We learn this over all-scales of the training image, by constructing an image pyramid with some scale factor (*e.g.*, $1/2$) between every two scales. Second, at inference time, we exploit the iterative sampling nature of the diffusion model to make possible jointly sampling frequency of all scales with consistency. Specifically, we use Laplacian decomposition to take only the corresponding frequency for each scale and from there we reconstruct an image. This is done at every iteration of the diffusion sampling. Last, we extend our sampling technique to work for sequence of images, by further incorporating the structure of the problem. For instance, our method makes sure that the sampled zoomed-in images are consistent with the corresponding regions of the original image. Overall, our contributions are several-folds:

- We explore the problem of extending single image prior to tasks of image sequence sampling, *e.g.*, zoom-in.
- We propose a simple learning framework for single image patch-level prior of all scales.
- We design and experimented with our sampling techniques to sample from our trained model that can achieve great single image unconditional sampling as well as joint and consistent sampling of an images sequence.

2 Related Work

Learning Single Image Prior. Several recent works have demonstrated that generative modeling can be used to capture the distribution of patches from a single image. SinGAN [1] and InGAN [5] were the first to propose this approach using generative adversarial networks (GANs) [6]. In particular, SinGAN shows results on unconditional sampling, where the patch statistics at every scale in the sampled image obey those of the training image. Furthermore, these methods can be applied to various vision tasks, such as super-resolution [1] and image retargeting [5], etc. Subsequently, researchers have explored other types of generative models to replace GANs, including the use of diffusion models in several recent works [8], [11], [12]. The works closest to ours are [8], [11], [12], as they also employ diffusion models for generative modeling. However, our method does not require conditional super-resolution modeling across scales of an image. Specifically, our method allows for sampling different scales of an image, as well as sampling a sequence of images, *all in parallel*. We also achieve high consistency across images, while each sample still retains faithful patch statistics at all scales. Most importantly, none of these works explore using learned single-image prior to sample a sequence of images, jointly and consistently.

Consistent Sampling From Diffusion Models. Diffusion models [13], [14], [15], [16], [17] are a type of generative model trained by denoising noise-perturbed data at different noise levels. During inference, samples are obtained by iteratively denoising from random noise. This iterative sampling nature is particularly favorable for jointly sampling consistent content, as exemplified by many existing works [18], [19], [20], [21], [22], [23]. MultiDiffusion [18] and its variants [20], [23] consistently denoise overlapping

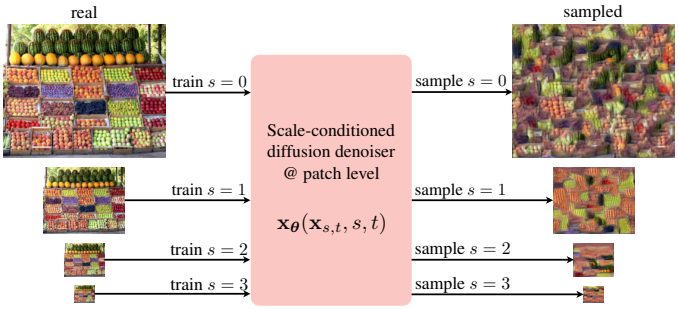

Fig. 2. Learning patch prior for multiple scales.

spatial regions of a large image canvas to achieve sampling mega-pixel images from a diffusion model trained on much smaller resolutions. DiffCollage [19] explores consistent sampling from general factor graphs, enabling more downstream applications with different structured probabilistic graphs. Generative Power of Ten demonstrates fascinating zoom-consistent sequence sampling from a large text-conditioned model, while [22] applies a similar idea to sample hybrid images perceived differently at varying viewing distances. Our sampling techniques are based on this line of work. In particular, we are largely inspired by Factorized Diffusions [22] for our single unconditional sampling and Generative Powers of Ten [24] for the zoom-in application. However, while these works explore diffusion priors trained on large internet-scale datasets, our diffusion prior is focused on a single image, at a patch level.

3 Method

3.1 Training Diffusion Models on Multi-scale Patch Distributions

Objective. Given a single image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, there is a collection of small $P \times P$ patches that forms a dataset of interest, $\mathcal{D} = \{\mathbf{p}^{(i)}\}_i$ with $\mathbf{p} \in \mathbb{R}^{P\times P\times C}$. This provides us with an empirical distribution $p_{\mathcal{D}}(\mathbf{p}) = \frac{1}{N}\sum_i \delta(\mathbf{p} - \mathbf{p}^{(i)})$. Our goal is that at inference time, a sampled image $\mathbf{x}_{\text{sample}}$ contains its collection of all $P \times P$ patches sampled from $p_{\boldsymbol{\theta}}(\mathbf{p}) \approx p_{\mathcal{D}}(\mathbf{p})$, where $p_{\boldsymbol{\theta}}(\mathbf{p})$ is a learned distribution parametrized by $\boldsymbol{\theta}$.

We want to capture patch statistics from multiple scales in order to unconditionally sample a single image. The intuition is that for the same patch size P, the small $P \times P$ patches at the coarse scale of the image represents the layout and structure, while the $P \times P$ patches at the fine scale of the image captures the textures. In order to unconditionally sample an image, we are required to capture all-scale patch distribution of the original image. Specifically, given a single image $\mathbf{x}$, we construct an image pyramid $\{\mathbf{x}_s\}_{s=0}^{N-1}$ with some relative scale factor ρ between two consecutive scales, *e.g.*, $\rho = 1/2$, where $\mathbf{x}_{s=0} = \mathbf{x}$. Each scale now has a corresponding patch distribution to be learned.

Diffusion models. We use Diffusion models [13], [14], [15], [16], [17], [25] for this generative learning task. In the plain formulation of diffusion models [25], the forward process adds different amount of noise $t \in [0, 1]$ to the clean signal $\mathbf{x} \sim p_{\mathcal{D}}(\mathbf{x})$ and create a noisy signal $\mathbf{x}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the variance-preserved schedules α_t,

σ_t are smooth functions chosen such that $\alpha_0 = \sigma_1 = 1$ and $\alpha_1 = \sigma_0 = 0$. Then, a neural network denoiser $\mathbf{x}_{\boldsymbol{\theta}}$ is trained to reconstruct the clean signal, given the noisy image and the scale of noise. Specifically, the training objective is

$$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}, t\sim\mathcal{U}(0,1), \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[w_t \|\mathbf{x}_{\boldsymbol{\theta}}(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, t) - \mathbf{x}\|_2^2\right], \quad (1)$$

where the w_t is some weight depends on the noise level t. At inference time, one first sample from the prior distribution (a gaussian noise) $\mathbf{x}_1$ and use the trained denoiser $\mathbf{x}_{\boldsymbol{\theta}}$ with an ODE/SDE sampler to iteratively get a clean sample $\mathbf{x}_0 = \mathbf{x}$.

Neural network architecture for denoiser $\mathbf{x}_{\boldsymbol{\theta}}$. To capture the distribution of patches, instead of explicitly gather patch datasets, we use a convolutional neural network of small $P \times P$ receptive field on an entire image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$. Specifically, our first convolutional layer has a kernel size $P \times P$ and all the later convolutional layers use a kernel size 1×1. We apply the forward process to the entire image and the objective used is to reconstruct the clean image with the objective in eq. (2). We claim that this achieves the objective stated, *i.e.*, if we use this trained denoiser to iteratively denoised an image canvas starting from noise, the collection of overlapping patches are sampled from the desired $P \times P$ patch distribution from the clean image $\mathbf{x}$.

Our diffusion training scheme. We adopt this described CNN as our denoiser for patch distribution learning. Since we learn from multi-scale dataset, our denoiser $\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_{s,t}, s, t)$ is then additionally conditioned on a scale signal $s \in \{0, \ldots, N-1\}$. In a training iteration, we randomly sample a pair of scale signal and the image at that scale $(s, \mathbf{x}_s)$ with $s \sim \text{Unif}\{0, \ldots, N-1\}$ for gradient decent. So the modified training objective is

$$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}, s\in\text{Unif}\{0,\ldots,N-1\}, t\sim\mathcal{U}(0,1), \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[w_t \|\mathbf{x}_{\boldsymbol{\theta}}(\alpha_t \mathbf{x}_s + \sigma_t \boldsymbol{\epsilon}, s, t) - \mathbf{x}_s\|_2^2\right]. \quad (2)$$

The full training algorithm is in algorithm 1. Our training algorithm is simple, where each scale condition captures the patch distribution at only the correponding scale. If one iteratively sample a image for each scale, respectively, it is expected that these images each only captures a certain scale patch distribution (see fig. 2). In the next section, we discuss how to sample a image with all-scales patch distributions.

Algorithm 1 Training

repeat
  $s \sim \text{Unif}\{0, 1, ..., N-1\}$
  $t \sim \mathcal{U}(0, 1)$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Take gradient descent step on
    $\nabla_{\boldsymbol{\theta}} w_t \|\mathbf{x}_{\boldsymbol{\theta}}(\alpha_t \mathbf{x}_s + \sigma_t \boldsymbol{\epsilon}, s, t) - \mathbf{x}_s\|_2^2$
until converged

3.2 Unconditional Single Image Sampling

We propose a sampling technique to communicate and seemlessly combine the patch distribution sampled from *all scales* at every sampling iteration, and thus the goal of unconditional image generation.

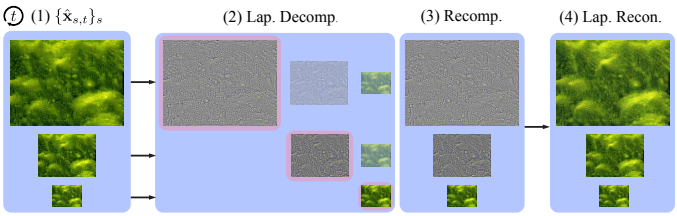

Fig. 3. The laplacian "re-composition" algorithm described in section 3.2, used in every iteration of diffusion sampling.

Commonly, with a denoiser $\mathbf{x}_{\boldsymbol{\theta}}$ trained, at some iteration of a diffusion sampler, *e.g.*, ddpm [14], starting at a noise level t and stepping by Δt, we have

$$\hat{\mathbf{x}}_t \leftarrow \mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \quad (3)$$

$$\mathbf{x}_{t-\Delta t} \leftarrow \text{DDPM-UPDATE}(\hat{\mathbf{x}}_t, \mathbf{x}_t, \Delta t) \quad (4)$$

where $\hat{\mathbf{x}}_t$ is the prediction of clean image $\mathbf{x}_0$ at noise level t, and the updated $\mathbf{x}_{t-\Delta t}$ is just an linear interpolation between $\hat{\mathbf{x}}_t$ and $\mathbf{x}_t$ with a small noise added as a result of ddpm update. Here, note that this $\hat{\mathbf{x}}_t$ in prediction space is agnostic to the diffusion parametrization and can be obtained in, *e.g.*, the noise $\boldsymbol{\epsilon}$-parametrization [14] and velocity $\mathbf{v}$-parametrization [26], via a simple re-parametrization calculation.

Considering that these $\hat{\mathbf{x}}_t$ predictions can be thought of as blurry approximations of the clean image $\mathbf{x}_0$, we can apply meaningful image operators as if we were working with clean images. Based on this insight, our Laplacian recomposition algorithm, at *every step* of diffusion sampling, performs the following steps (illustrated in fig. 3): (1) a prediction space pyramid $\{\mathbf{x}_{s,t}\}_{s=0}^{N-1}$ is given by querying the denoiser individually for each scale from the last iteration, (2) decompose each scale into a Laplacian pyramid [27], (3) extract only the highest-frequency Laplacian layer from each scale (except for the coarsest scale, which retains the original image), and use these to recompose a new Laplacian pyramid, and (4) finally reconstruct the pyramid to produce the updated predicted pyramid $\{\mathbf{x}_{s,t}\}_{s=0}^{N-1}$. Subsequently, the DDPM update is performed on this updated pyramid for each scale individually. Intuitively, this algorithm isolates the highest frequencies captured at each scale (those not captured by coarser scales) and combines them using a straightforward Laplacian reconstruction. The full sampling algorithm is provided in algorithm 2.

Algorithm 2 Unconditional Single Image Sampling

$\{\mathbf{x}_{s,t=1}\}_{s=0}^{N-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ initial a noise pyramid
$\{\hat{\mathbf{x}}_{s,t=1}\}_{s=0}^{N-1} = \{\mathbf{0}\}_{s=0}^{N-1}$ ▷ initial "predicted" pyramid
for t in finite step sampling schedule **do**
  for $s = 0, \ldots, N-1$ **do**
    $\hat{\mathbf{x}}_{s,t} = \mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_{s,t}, s, t)$
  end for
  $\{\hat{\mathbf{x}}_{s,t}\}_{s=0}^{N-1} = \text{LAPLACIAN-RECOMP}(\{\hat{\mathbf{x}}_{s,t}\}_{s=0}^{N-1})$
  for $s = 0, \ldots, N-1$ **do**
    $\mathbf{x}_{s,t} = \text{DDPM-UPDATE}(\hat{\mathbf{x}}_{s,t}, \mathbf{x}_{s,t}, t)$
  end for
end for

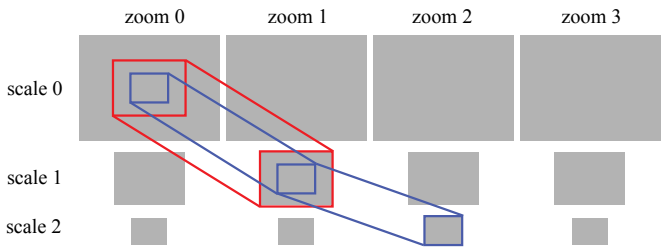

Fig. 4. Illustration of the zoom-scale space with diagonals consistency.

3.3 Zoom-in Sequence Sampling

Our formulation can also easily extend to sample a sequence of images. We show how to sample a zoom-in sequence of images. Suppose we are dealing we zoom-in by some factor every time, we denote the sequence of M zoom levels as $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(M-1)}$. Note that by definition, these zoom levels should have a consistency across different scales of different zoom levels. For example, the central crop of the scale 0 of the zoom 0 should be consistent with the scale 1 of zoom 1, and so on. And this zoom consistency can be illustrated in a zoom-scale space, where all the diagonals has a chain of consistency (see illustration in fig. 4).

We will enforce this consistency by simple averaging the crops on the diagonal and replace them with the averaged results in the $\hat{\mathbf{x}}$-prediction space, and we apply this at every iteration. During each iteration, this consistency constrain is imposed first, then the laplacian recomposition for each zoom level, respectively.

4 Experiments

4.1 Unconditional single image generation

We unconditionally sample individual images from the single image prior according to the technique described in section 3.2. Specifically, we set the relative factor between two scales to $\rho = 1/2$ and we trained the single image prior on mostly images of resolution 250 pixels on the longer side. We construct a pyramid for 4 scales and set the receptive field to be $P = 11$ so that the coarsest scale is around twice as large as the receptive field. We set batch size to be 16 over $t \sim \mathcal{U}(0, 1)$ once a scale s is chosen. Our denoiser comprises of around 4×10^6 parameters, We train the model for 3×10^5 steps, which takes around 12 hours on a A6000 gpu.

Results. We show results for several images in fig. 5. These sampled images are uncurated and share patch distributions of the original images. However, it lacks some high frequency details and contain some regions of blurriness. We also compare ours with a baseline method [8] in fig. 6. And in terms of the visual fiedlity the baseline seems to perform better. However, we expect our algorithm can work better with some other improvements. It is very excited to see for the first time a method not using conditional super-resolution supervision during the training can work almost on par with those using, and our training algorithm is much easier not modifying the original diffusion models training at all.

4.2 Zoom-in sequence generation

We use the algorithm specified in section 3.3. The model is trained the same way as specified in section 4.1 Specifically, we choose the zoom factor to be the inverse of the relative factor $1/\rho = 2$. We test our method on fractal-like photographs and expect to get every zoom level looks like the original image and with the perfect zoom-consistency.

Results. Our sampled zoom sequence, shown in fig. 7 though have perfect zoom-consistency due to the nature of the sampling algorithm, each zoom level does not look like the original image in terms of frequency spectrums.

5 Conclusion

We propose a newly proposed training-sampling techinique for diffusion models that is trained on a single image. We also enable image sequence sampling such as zoom-in sqeuence. We experimentally found that even without the conditional super-resolution supervision, we could still do unconditional image generation with high fidelity. We also experimented on the zoom-in sequence sampling with a sampling algorithm proposed.

Limitations. Our single image unconditional generation results are not as good as some other baseline methods. Our zoom-in sequence sampled are of low fidelity.

Future works. We will improve our single image unconditional generation quality as well as the zoom-in sequence sampling. Primarily, we need to study why some high frequency details are not captured and why regions of blurs occurs, which almost looks like some disagreements of overlapping patch generations. We will also explore other consistent sampling task and not restricted to only "sequence", but any collection of contents, *e.g.*, trained two single image diffusion priors and mixing the two together.

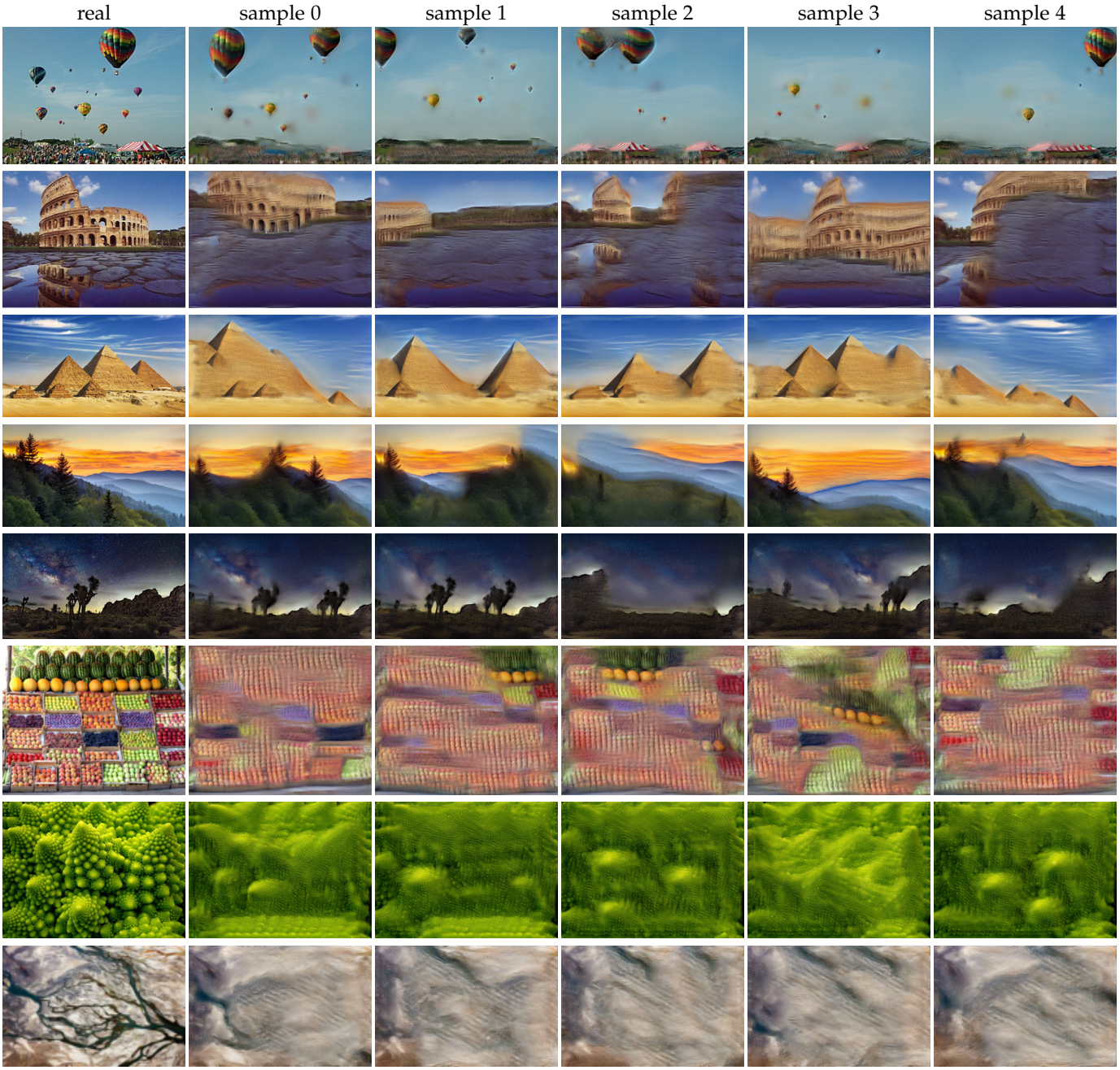

Fig. 5. Comparison between real images and sampled images uses the consistent sampling technique described in **??**.

REFERENCES

[1] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a Generative Model from a Single Natural Image," Sep. 2019, arXiv:1905.01164 [cs]. [Online]. Available: http://arxiv.org/abs/1905.01164

[2] A. Buades, B. Coll, and J.-M. Morel, "A Non-Local Algorithm for Image Denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. San Diego, CA, USA: IEEE, 2005, pp. 60–65. [Online]. Available: http://ieeexplore.ieee.org/document/1467423/

[3] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *2009 IEEE 12th International Conference on Computer Vision*. Kyoto: IEEE, Sep. 2009, pp. 349–356. [Online]. Available: http://ieeexplore.ieee.org/document/5459271/

[4] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *CVPR 2011*, Jun. 2011, pp. 977–984, iSSN: 1063-6919.

[5] A. Shocher, S. Bagon, P. Isola, and M. Irani, "InGAN: Capturing and Retargeting the "DNA" of a Natural Image," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 4491–4500. [Online]. Available: https://ieeexplore.ieee.org/document/9009560/

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[7] N. Granot, B. Feinstein, A. Shocher, S. Bagon, and M. Irani, "Drop the GAN: In Defense of Patches Nearest Neighbors as Single Image Generative Models," Mar. 2021, arXiv:2103.15545 [cs] version: 1. [Online]. Available: http://arxiv.org/abs/2103.15545

[8] V. Kulikov, S. Yadin, M. Kleiner, and T. Michaeli, "SinDDM: A Single Image Denoising Diffusion Model," Dec. 2022, arXiv:2211.16582 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2211.16582

[9] T. Hinz, M. Fisher, O. Wang, and S. Wermter, "ConSinGAN - Improved Techniques for Training Single-Image GANs," Nov. 2020, arXiv:2003.11512 [cs]. [Online]. Available: http:

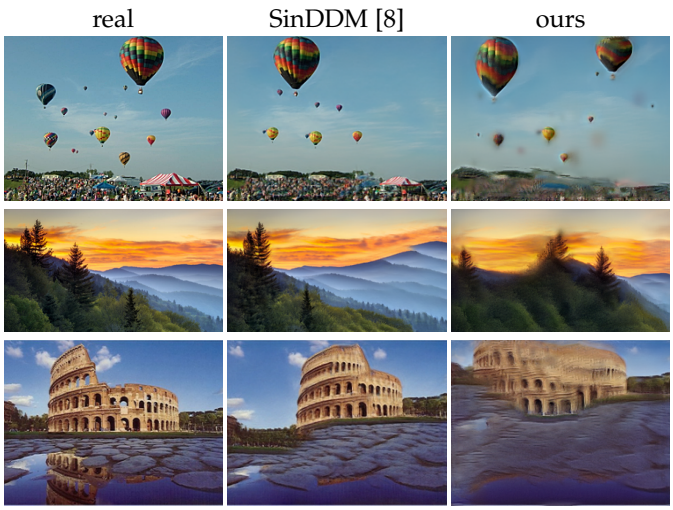

Fig. 6. Comparison between real v.s. SinDDM and ours, SinDDM performs better in visual experience.

//arxiv.org/abs/2003.11512

[10] Z. Zheng, J. Xie, and P. Li, "Patchwise Generative ConvNet: Training Energy-Based Models from a Single Natural Image for Internal Learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 2960–2969. [Online]. Available: https://ieeexplore.ieee.org/document/9577881/

[11] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "SinDiffusion: Learning a Diffusion Model from a Single Natural Image," Nov. 2022, arXiv:2211.12445 [cs]. [Online]. Available: http://arxiv.org/abs/2211.12445

[12] Y. Nikankin, N. Haim, and M. Irani, "SinFusion: Training Diffusion Models on a Single Image or Video," Jun. 2023, arXiv:2211.11743 [cs]. [Online]. Available: http://arxiv.org/abs/2211.11743

[13] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," Nov. 2015, arXiv:1503.03585 [cond-mat, q-bio, stat]. [Online]. Available: http://arxiv.org/abs/1503.03585

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," Dec. 2020, arXiv:2006.11239 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2006.11239

[15] Y. Song and S. Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution," Oct. 2020, arXiv:1907.05600 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1907.05600

[16] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations," Feb. 2021, arXiv:2011.13456 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2011.13456

[17] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the Design Space of Diffusion-Based Generative Models," Oct. 2022, arXiv:2206.00364 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2206.00364

[18] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation," Feb. 2023, arXiv:2302.08113 [cs]. [Online]. Available: http://arxiv.org/abs/2302.08113

[19] Q. Zhang, J. Song, X. Huang, Y. Chen, and M.-Y. Liu, "DiffCollage: Parallel Generation of Large Content with Diffusion Models," Mar. 2023, arXiv:2303.17076 [cs]. [Online]. Available: http://arxiv.org/abs/2303.17076

[20] Y. Lee, K. Kim, H. Kim, and M. Sung, "SyncDiffusion: Coherent Montage via Synchronized Joint Diffusions," Oct. 2023, arXiv:2306.05178 [cs]. [Online]. Available: http://arxiv.org/abs/2306.05178

[21] D. Geng, I. Park, and A. Owens, "Visual anagrams: Generating multi-view optical illusions with diffusion models," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [Online]. Available: https://arxiv.org/abs/2311.17919

[22] ——, "Factorized Diffusion: Perceptual Illusions by Noise Decomposition," Apr. 2024, arXiv:2404.11615 [cs]. [Online]. Available: http://arxiv.org/abs/2404.11615

[23] S. Frolov, B. B. Moser, and A. Dengel, "Spotdiffusion: A fast approach for seamless panorama generation over time," 2024. [Online]. Available: https://arxiv.org/abs/2407.15507

[24] X. Wang, J. Kontkanen, B. Curless, S. Seitz, I. Kemelmacher, B. Mildenhall, P. Srinivasan, D. Verbin, and A. Holynski, "Generative Powers of Ten," Dec. 2023, arXiv:2312.02149 [cs]. [Online]. Available: http://arxiv.org/abs/2312.02149

[25] D. P. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2107.00630

[26] M. Mardani, J. Song, J. Kautz, and A. Vahdat, "A Variational Perspective on Solving Inverse Problems with Diffusion Models," Sep. 2023, arXiv:2305.04391 [cs, math, stat]. [Online]. Available: http://arxiv.org/abs/2305.04391

[27] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, 1983.

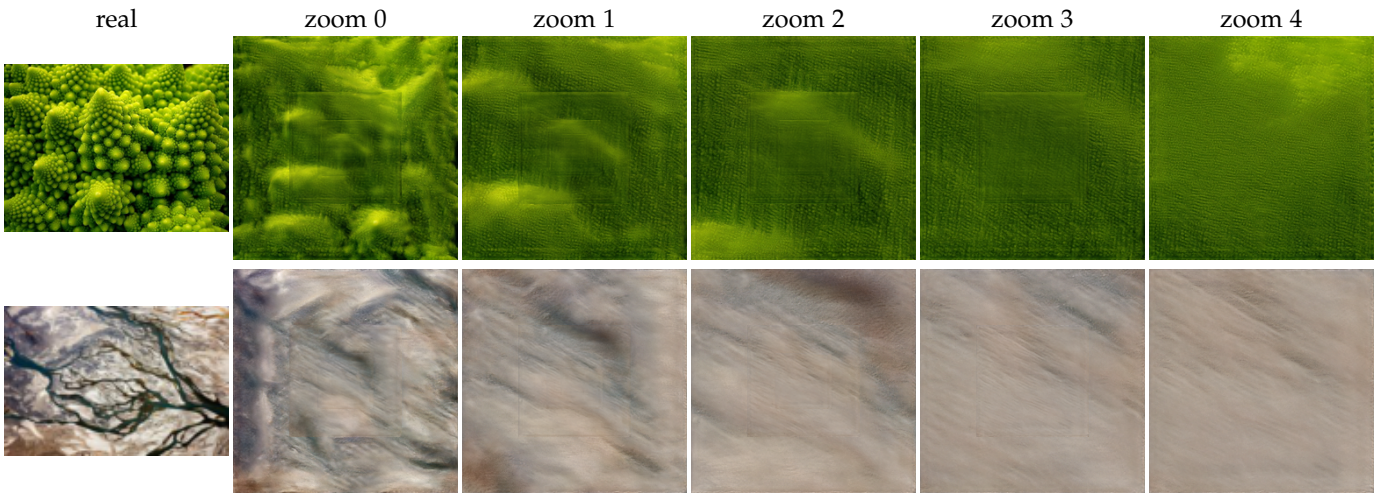

Fig. 7. Zoom in sequence sampling with 5 zoom levels for two fractal-like photographs.