Consistent Image Collection Sampling from Single Image Diffusion Prior
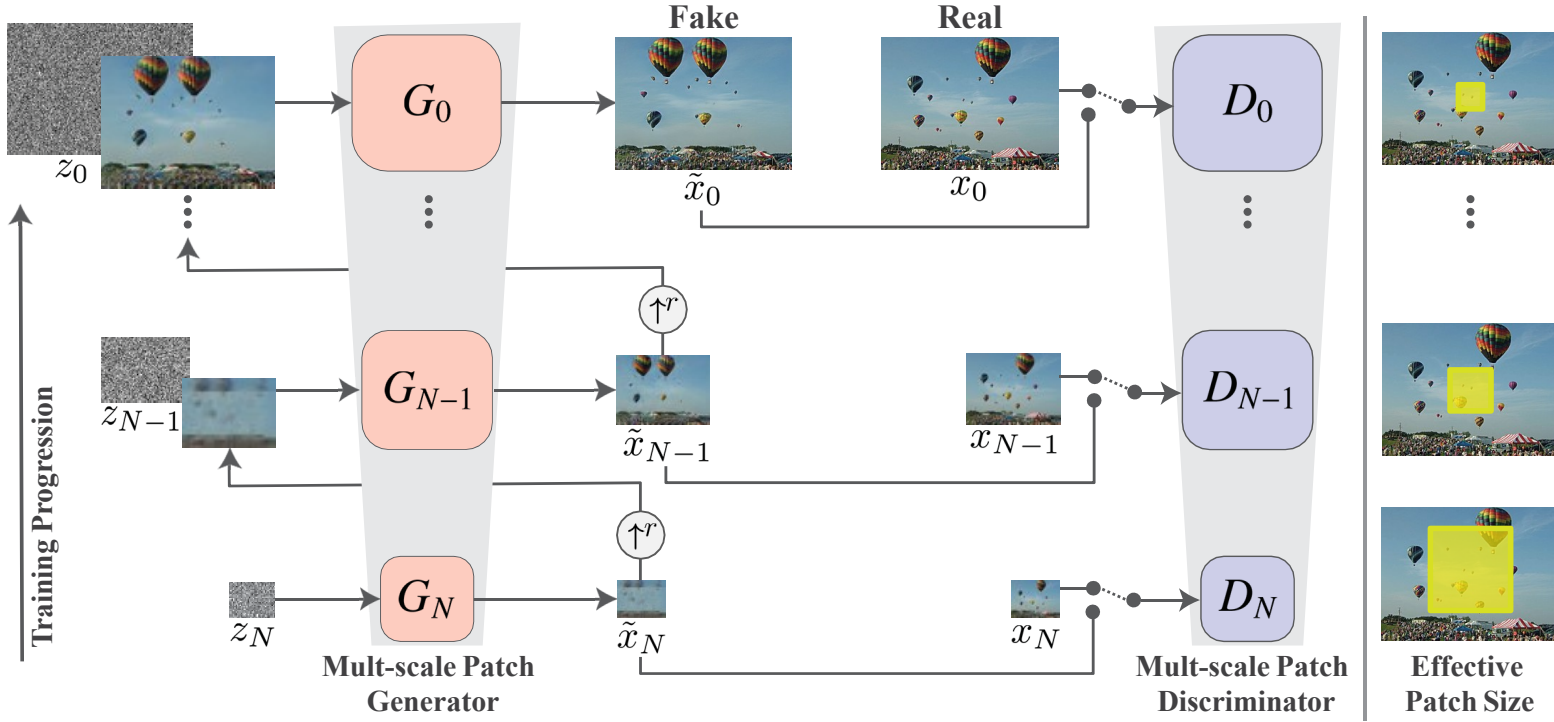
Haojun Qiu

Department of Computer Science, University of Toronto

## Prior Works

Single Image Unconditional Synthesis.
SinGAN [1], SinDDM[2], difficult to extend conditional SR for task like infinitely zoom-in

Consistent Content Sampling from Internet-scale Diffusion Prior.
- MultiDiffusion [3], average the score prediction for spatial consistency
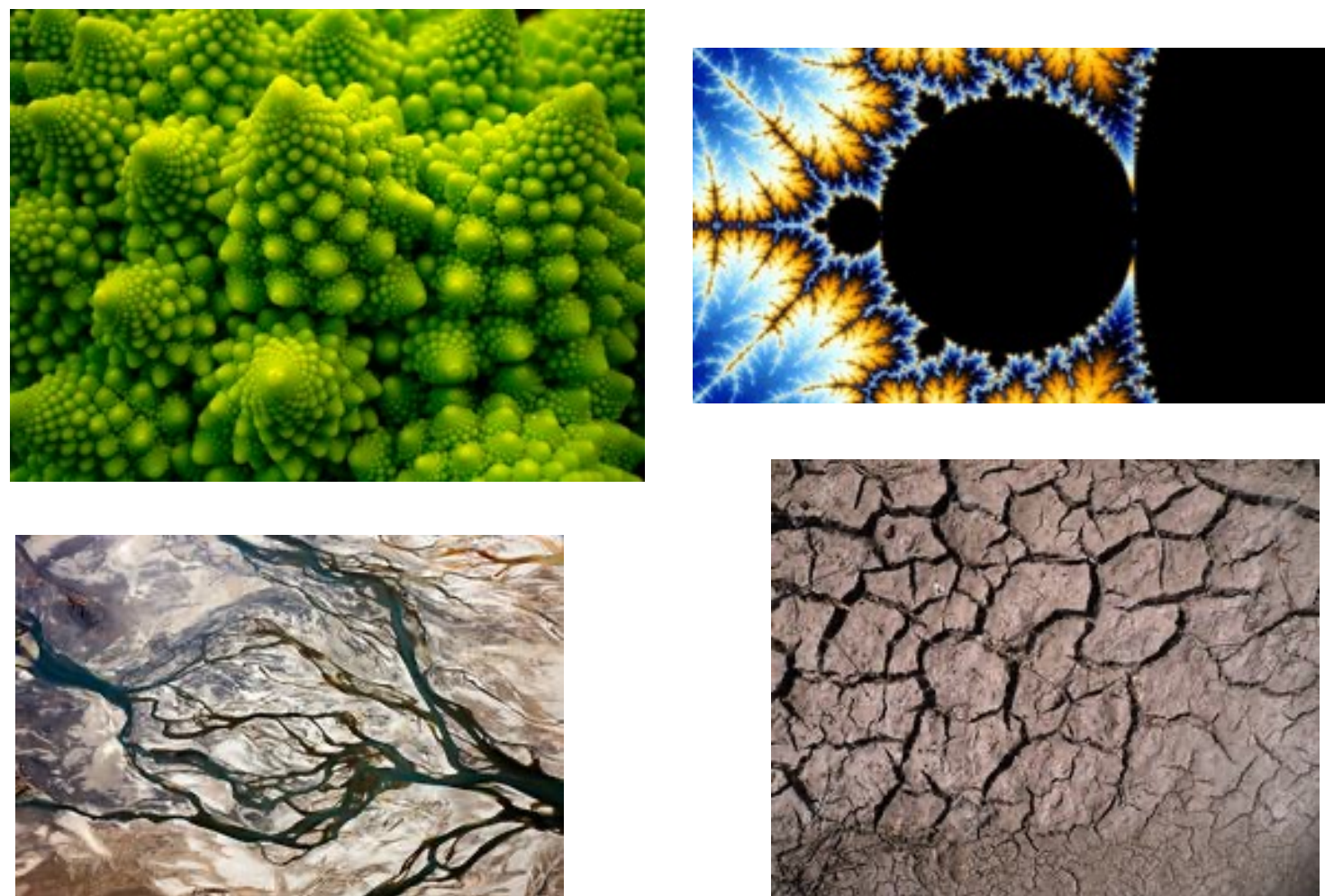- Generative Power of Ten [4], Factorized Diffusion [5] both applying consistency over scales space.
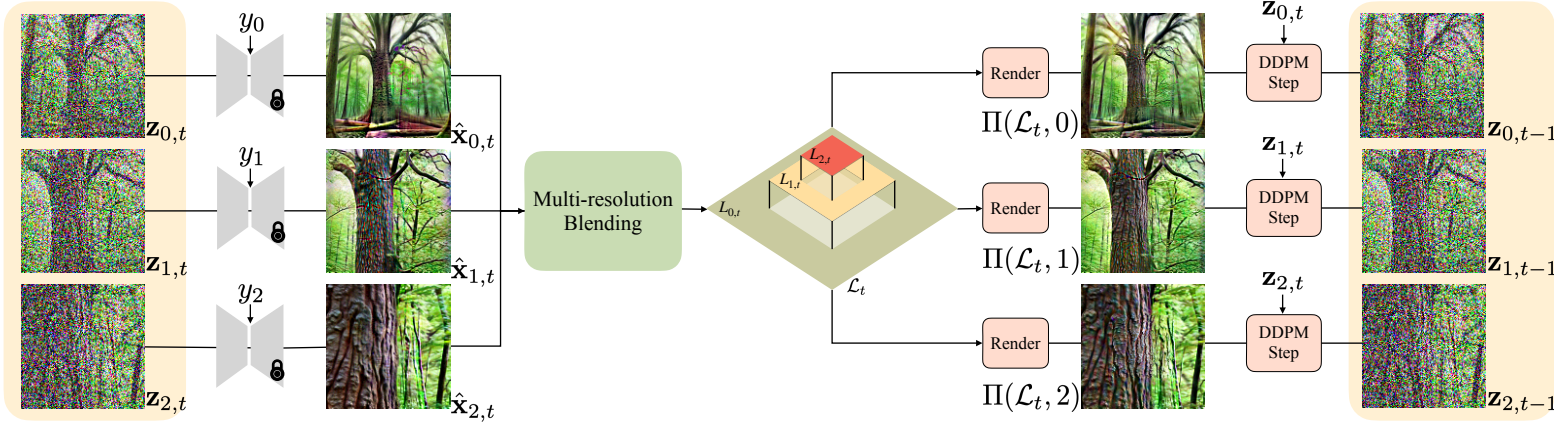
We study this for single image diffusion prior.

Motivation & Research Question

How can we sample a *collection* of images, *jointly and consistently*, from a *single-image prior*? Concrete example include infinitely zoom-in as a sequence of images.


- Diffusion for its easiness applying consistency
- A new training-sampling scheme without conditional SR, possibilities for more tasks
- Each task is just a sampling algorithm

References

[1] Shahams et al., SinGAN, ICCV, 2019
[2] Kulikov et al., SinDDM, ICML, 2023
[3] Bar-Tal et al., MultiDiffusion, ICML, 2023
[4] Wang et al., Generative Powers of Ten, CVPR, 2024
[5] Geng et al., Factorized Diffusion, ECCV, 2024

Methods

Training. Multiscale with small receptive field CNN with Diffusion objective.

Scale-conditioned diffusion denoiser @ patch level

$$\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_t, s, t)$$
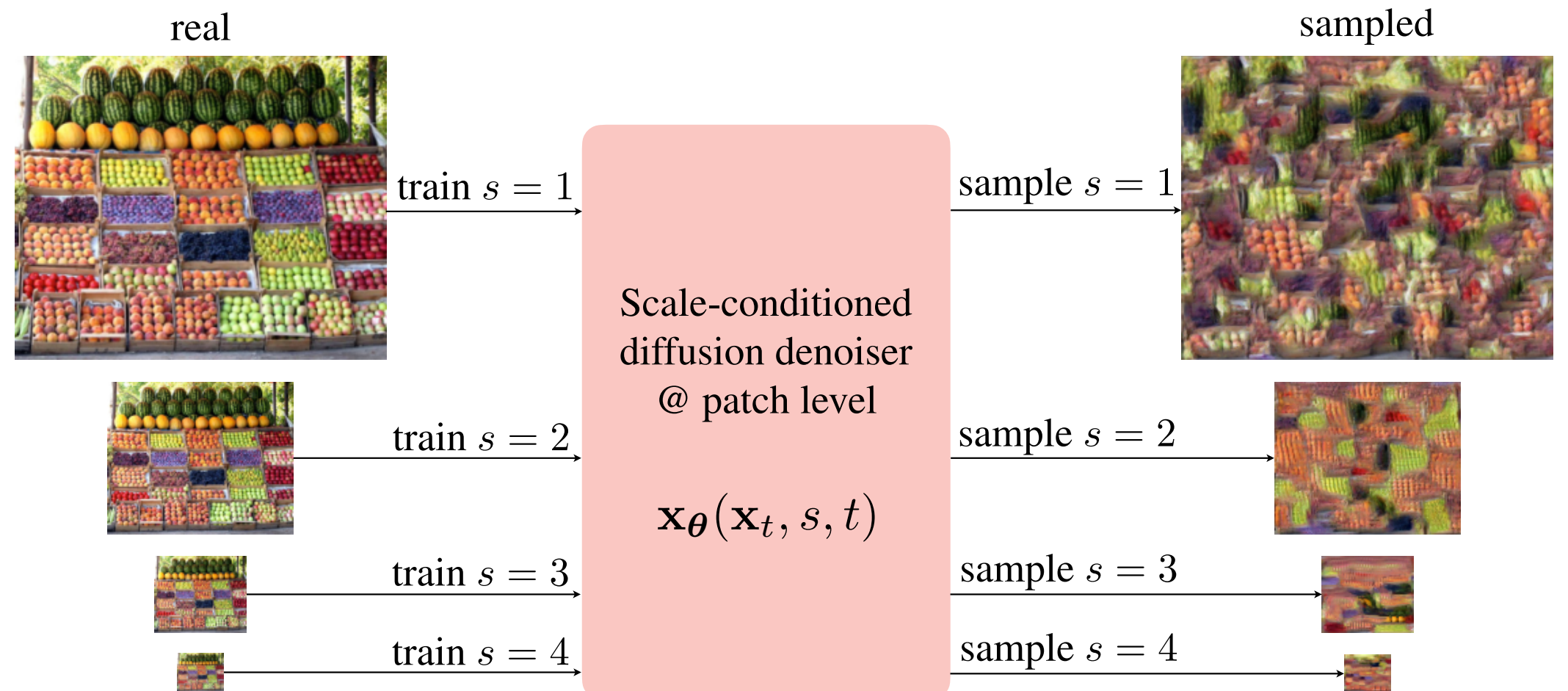
Single Image Unconditional Sampling. Iteratively done "laplacian recomposition".

$\{\hat{\mathbf{x}}_{t,s}\}_s$ Laplacian Decomp.   Laplacian Recomp.   Laplacian Recon.
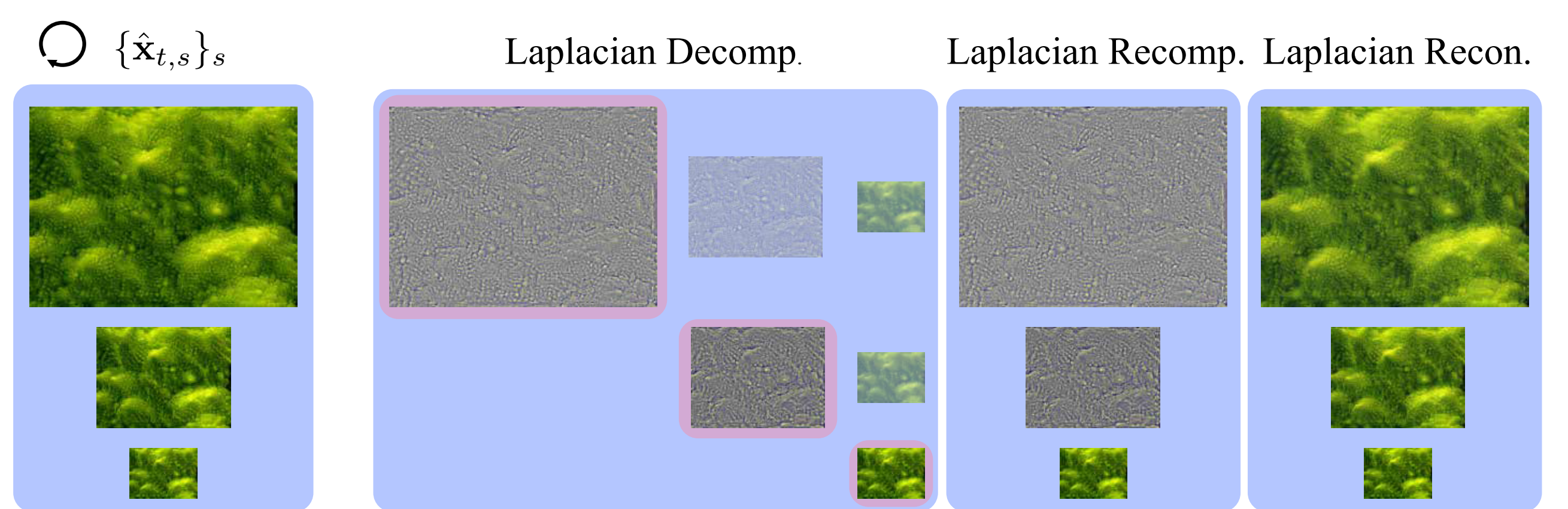
Zoom-in. Iteratively score averaging on the diagonals of the "zoom-scale" matrix

zoom 0   zoom 1   zoom 2   zoom 3

scale 0

scale 1

scale 2
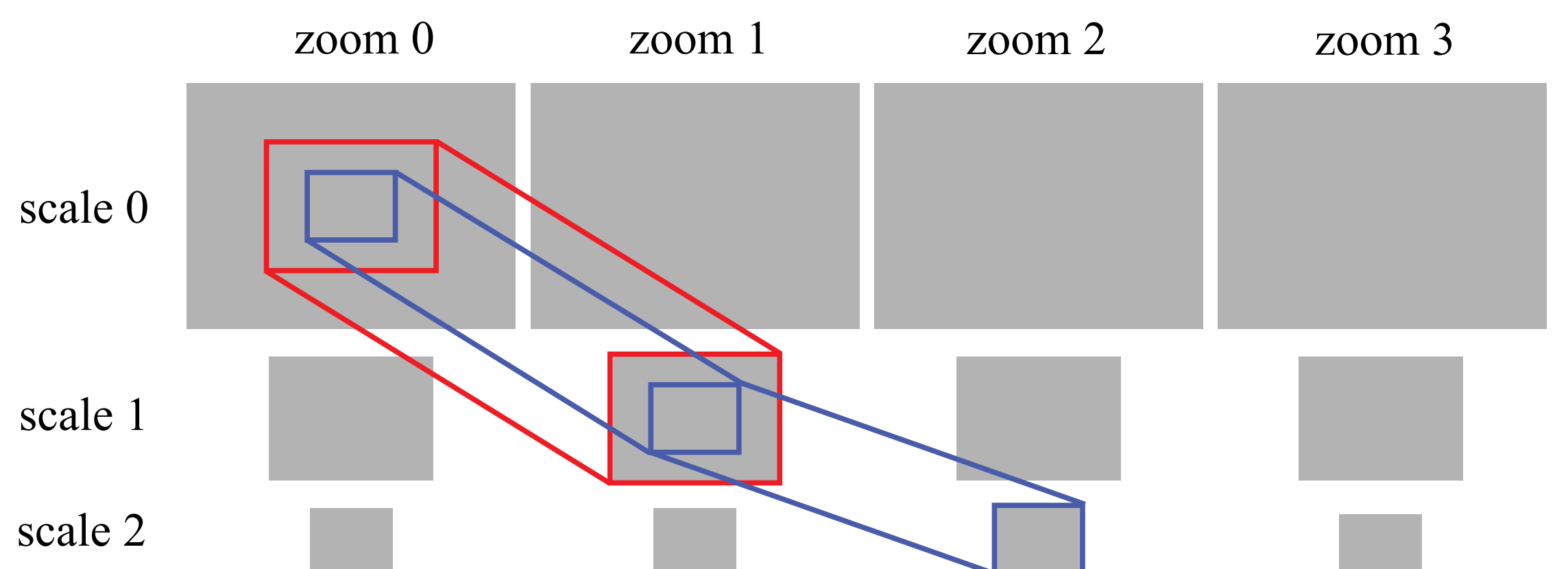
Experimental Results

Single Image Unconditional Sample. Some fine results, not yet as good as conditional SR baselines. Could be (1) distribution shift, (2) Training, e.g., NN is not exactly doing overlapping patch samples, (3) Calibration in pyramid building, (4) inference time sampling sub-optimal

original   samples
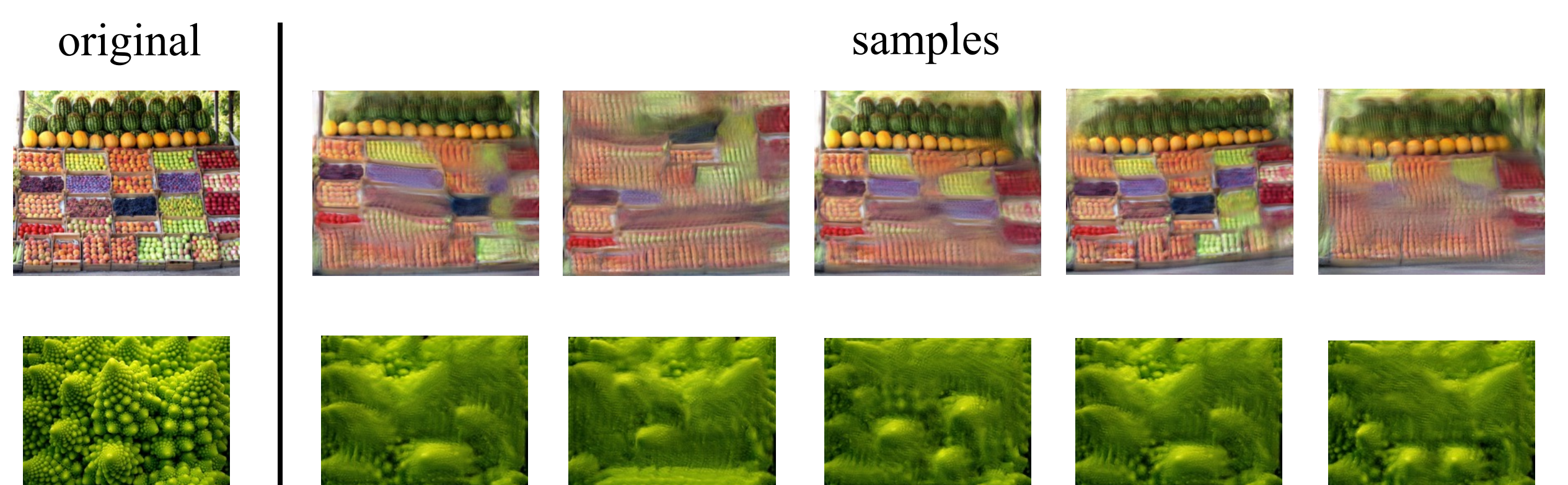
Zoom-in. Lack frequency details, partially due to sub-optimality in single image unconditional sampling, additionally many aspects of the algorithm could be improved,

zoom 0   zoom 1   zoom 2   zoom 3   zoom 4