Learning Generative ConvNets via Multi-grid Modeling and Sampling

Ruiqi Gao[1]*, Yang Lu[2]*, Junpei Zhou[3], Song-Chun Zhu[1], Ying Nian Wu[1]
[1] University of California, Los Angeles, USA [2] Amazon, [3] Zhejiang University, China
ruiqigao@ucla.edu, ylumzn@amazon.com
jpzhou1996@gmail.com, {sczhu, ywu}@stat.ucla.edu

Abstract

The supplementary materials include more results on synthesis and diagnosis.

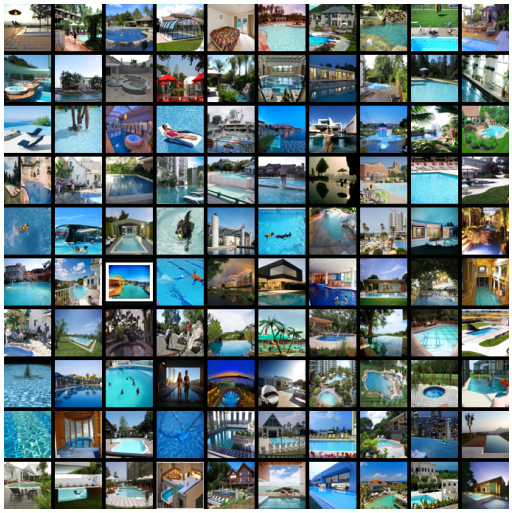
1. Synthesis

Figure 1 shows synthesized results learned on Street View Housing Numbers (SVHN) dataset. Figures 2-6 show synthesized results learned on several categories of MIT place205 dataset. The number of training images is 7,300 for swimming pool category and 15,100 for other categories.

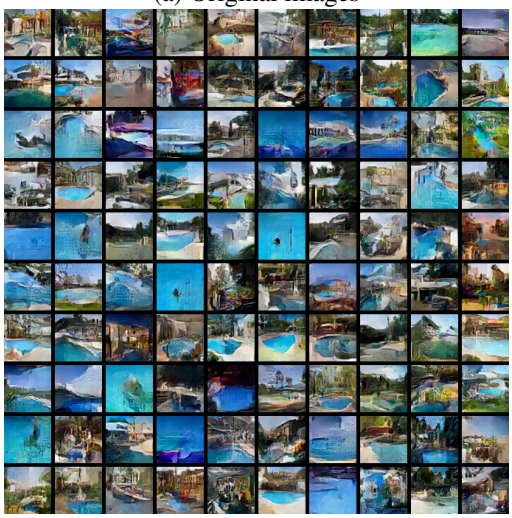
Observed DCGAN

Single-grid Multi-grid CD

Figure 1. Synthesized images by models learned from the SVHN dataset. CD1 and persistent CD cannot synthesize realistic images and their results are not shown.



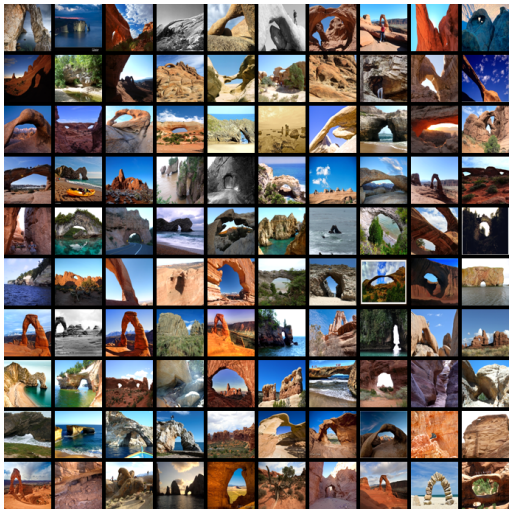
(a) Original images



(b) Synthesized images

Figure 2. Synthesized images generated by multi-grid learned on `swimming pool` category in MIT places205

*Equal contributions.



(a) Original images

(b) Synthesized images

Figure 3. Synthesized images generated by multi-grid learned on rock category in MIT places205

(a) Original images

(b) Synthesized images

Figure 4. Synthesized images generated by multi-grid learned on building facade category in MIT places205

2. Diagnosis

To monitor model fitting and synthesis, we calculate the values of scoring function $f_\theta(Y)$ after training. Table 1 shows the results after 400 iterations of training on CelebA dataset. We randomly sample 10,000 images that are not included in the training dataset from CelebA for testing, and use images randomly sampled from MIT places205 as negative examples. Compared with negative images, scores of training and testing images are higher and close to each other. Scores of training and synthesized images are also close, indicating that the synthesized images are close to fair samples.

To monitor the stability of the multi-grid method, Fig. 7 shows the l_1-norm of the gradients over iterations in the experiment of learning from the CelebA dataset. The plots show that the learning is stable.

Table 1. Average $\pm$ standard deviation of the score $f_\theta(Y)$.

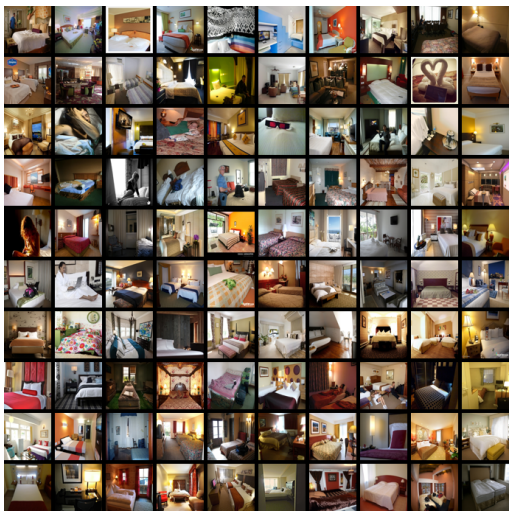
Images	grid1 ($\times 10^4$)	grid2 ($\times 10^5$)	grid3 ($\times 10^6$)
Training	5.33 ± 0.91	8.59 ± 1.12	2.59 ± 0.10
Testing	5.33 ± 0.89	8.27 ± 1.01	2.47 ± 0.10
Synthesized	5.15 ± 0.91	8.38 ± 1.17	2.58 ± 0.11
Negative	4.10 ± 1.00	5.42 ± 1.19	1.99 ± 0.11

To check that the learned model is not just memorizing training datasets, in fig. 8, we show some synthesized images, the corresponding observed images from which the initial 1×1 patches are down-sampled, and three nearest neighbors (in Euclidean distance) of the synthesized images in the training datasets.

(a) Original images

(b) Synthesized images

Figure 5. Synthesized images generated by multi-grid learned on forest road category in MIT places205



(a) Original images

(b) Synthesized images

Figure 6. Synthesized images generated by multi-grid learned on hotel room category in MIT places205

Besides Langevin dynamics, we also try to use Hamiltonian Monte Carlo (HMC) as sampler, and compare the efficiency of these two samplers. Our sampling method runs independent parallel chains, starting from 1×1 image and going through 30 sampling steps at each grid. For each grid, let Y_0 be the starting image and Y_{30} be the image after 30 sampling steps. We calculate the average pixel-wise correlation between Y_0 and Y_{30} across a batch of independent chains. For HMC sampler, we use 5 steps of leapfrog in each iteration. Results on CelebA dataset are shown in Table 2. Coarser grid has smaller correlation than finer grid, indicating that coarser grid creates variability, while finer grid provides refinement. HMC sampler has smaller correlations.

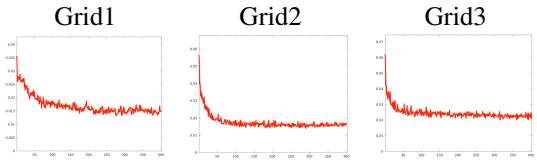


Figure 7. l_1-norm of gradients over iterations of learning from the CelebA dataset.

Table 2. Correlation between Y_0 and Y_{30} on CelebA dataset.

Langevin Dynamics			HMC		
grid 1	grid 2	grid 3	grid 1	grid 2	grid 3
0.62	0.83	0.94	0.62	0.82	0.91

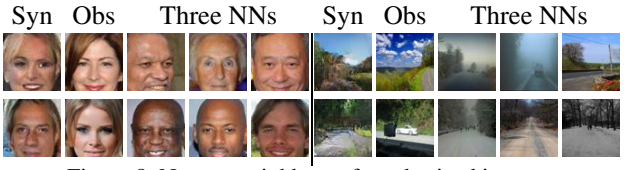


Figure 8. Nearest neighbors of synthesized images.