Edge-Selective Super-Resolution using SinGAN

Apurva Bhargava (ab8687), Nikhil Supekar (ns4486)
DS-GA 3001.004, Introduction to Computer Vision, Spring 2021

May 10, 2020

1 Introduction

Super-Resolution is the process of creating high-resolution images from low-resolution images. A super-resolution model improves or adds details within an image. The goal of single-image super-resolution (SISR) algorithms is to recover the details using the prior knowledge gained from a single image and/ or some heuristics. A low resolution (LR) image is the input to such an algorithm. The same image, upscaled to a higher resolution (HR), is the output. This is an under-determined inverse problem because the input LR image does not contain the full HR image information. The missing information is crucial in making the HR image look sharp (Rouf et al. (2016)). Super-resolution is useful for many image processing tasks with direct applications in medical imaging, face recognition, satellite imaging and surveillance (Nazeri et al. (2019)).

Related Work: SISR algorithms can be categorized based on their tasks (domain-specific and generic) or methodology (deterministic and learned). Domain-specific SISR algorithms focus on specific classes of images such as faces, scenes and graphics artwork, for example, Discriminative Generative Networks (Yu et al. (2016)), Progressive Adversarial Network (Wong et al. (2020)) and Face Super Resolution Net (Chen et al. (2018)). Generic SISR algorithms are developed for all kinds of images where the priors are typically based on primitive image properties such as edges and segments. These include interpolation-based methods (bicubic from Keys (1981) and Lanczos from Duchon (1979), which are speedy but lack accuracy), edge-based (Edge-informed SISR from Nazeri et al. (2019) and Fast Edge-directed SISR Rouf et al. (2016)), and patch-based methods (Yang et al. (2014)). The deep-learning based methods include Very Deep ConvNets (Kim et al. (2016)), EnhanceNet (Sajjadi et al. (2017)), Explicit Natural Manifold Discrimination (Soh et al. (2019)) and SinGAN (Shaham et al. (2019)).

2 Problem Definition

Single image super-resolution is challenging because high-frequency image content typically cannot be recovered from a single low-resolution image. It can be likened to a down-sampled, low-pass filtered version of our expected, super-resolved output. Without high-frequency information, the quality of the output image is limited. Further, this is an ill-posed problem since multiple HR images can correspond to one LR image (Soh et al. (2019)).

Super-resolution involves three tasks- upsampling, deconvolution, and denoising (Rouf et al. (2016)). The last of these tasks essentially smooths out most elements of an image except edges. Making super-resolution limited to the edges offers a way of generating high quality images without expending too much compute power. The single-image SR problem can be reformulated as an edge-deblurring or edge-in-painting problem, which makes it more well-posed. As discussed in the next section, we take the encodings learnt from variable-sized patches through a single-image GAN and then use MLP approximator to build a deterministic function that takes in pixel coordinates and outputs RGB (color) value for that pixel, making our technique somewhat independent of the smooth contour prior.

3 Proposed Approach

We propose 'edge-selective super-resolution' by building an MLP-function approximator, inspired by Jiang et al. (2020), on top of SinGAN (Shaham et al. (2019)) for super-resolving edges. Arbitrary querying of pixel values can also be thought of as 'infinite-resolution', delivering as high a quality as desired at the edges. We

achieved the model using a classical computer vision technique for edge detection (Canny-Edge detector) and multi-layered deep learning models for generating a super-resolution function.

3.1 SinGAN

SinGAN or Single Image Generative Adversarial Networks from (Shaham et al. (2019)) generate high-quality, realistic and natural random samples learned from a single natural image. SinGAN contains a pyramid of fully-convolutional GANs. Each GAN is responsible for learning the patch distribution at a different scale of the image. This allows generating new high quality, diverse samples which contain the same visual content as the image. The samples can be of arbitrary size and aspect ratio, have significant variability, and yet maintain both the global structure and the fine textures of the training image. SinGAN is not limited to texture images (i.e., it works for all images, natural or otherwise), and is not conditional (i.e. it generates samples from noise). Applications include super-resolution, paint-to-image, seamless editing, single image animation, etc.

Architecture:

- As shown in Figure 1, the input to generator G_N is a random noise image, z_N, and the generated image from the previous scale x_N, upsampled to the current resolution (except for the coarsest level which is purely generative).
- As the inputs for a level are upscaled, the effective patch size decreases as we go up the pyramid (marked in yellow in Figure 1). Both training and inference are done in a coarse-to-fine fashion.
- Each of the generators G_N, is coupled with a Markovian discriminator D_N, that classifies each of the overlapping patches of its input as real or fake.
- Training: At a given scale (or level N), generator G_N requires generating samples whose overlapping patches cannot be distinguished from the patches in the down-sampled training image, by the discriminator D_N. The objective function is given by

$$\min_{G_n} \max_{D_n} L_{adv}(G_n, D_n) + \alpha L_{rec}(G_n)$$

  The adversarial loss L_{adv} penalizes for the distance between the distribution of patches in x_n and the distribution of patches in generated samples. The reconstruction loss L_{rec} insures the existence of a specific set of noise maps that can produce x_n, which is an important feature for image manipulation.
- The generation process at level n involves all generators and all noise maps (noise images) up to this level. (Shaham et al. (2019)).

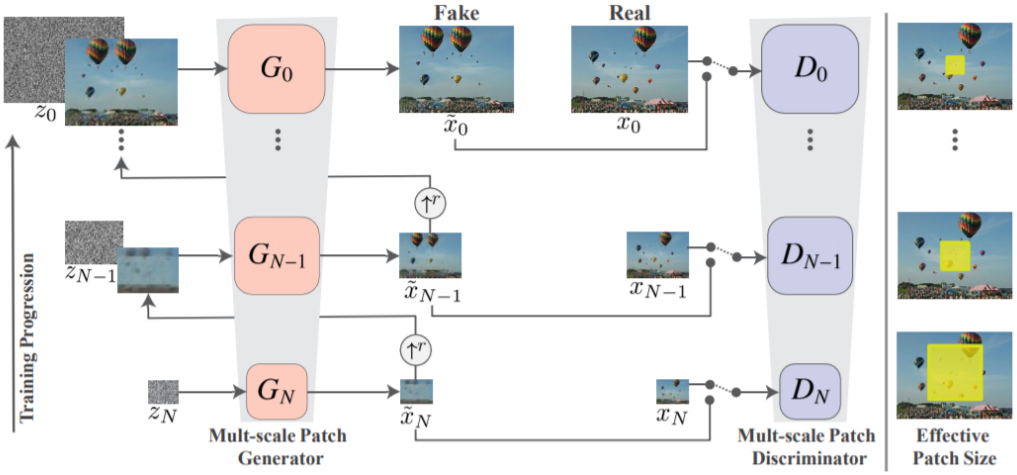


Figure 1: SinGAN Architecture: Multi-scale Pipeline (from Shaham et al. (2019))

3.2 Super Resolution as a function

Our key insight in extending SinGAN for super-resolution is to view an image as a function. Computers store a digitize version of images, making it a discrete function of pixel co-ordinates and output colors. However, real-world image have an underlying continuous distribution of various features that have much finer details about the image. In theory, if we were to have access to this function, we could look at the finest of details at any resolution that we please. Inspired by this view, we aim to learn the image as a continuous function of real co-ordinates instead of integer co-ordinates. With an access to such a function, we can view super-resolution as querying this function at arbitrary real co-ordinates between neighboring integer pixel co-ordinates, which can also be thought of as 'infinite' resolution.

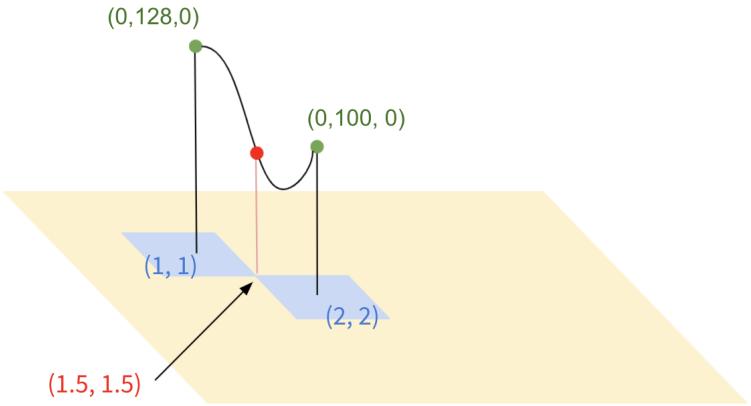


Figure 2: Image as a continuous function. (Example) 2x super-resolution queries the learned function for midpoints of integer co-ordinates.

3.3 Encoder-Decoder Architecture

We aid the learning of the continuous function by extracting image encodings from SinGAN. The hypothesis is that multi-scale generators from SinGN are capable of capturing higher-level features at lowest level and finer details at the highest level. Moreover, this architecture lends itself well to the task of super-resolution since it naturally upscales images at every level, allowing us to train for decimal co-ordinates, with a coordinate system set up on the lowest level. However, the original architecture does not have inherent image encodings in the generators. Therefore, we substitute the SinGAN generators with encoder-decoder style autoencoders to learn image encodings.

3.4 MLP Function Approximator

Most deep-learning models usually process a full image and output a complete image without any discriminative selection between its various elements. A simplification of the problem of applying super-resolution only to the edges, while using the information learned from an image, is to create a function that takes in real-valued pixel coordinates and outputs the required RGB value. Since neural networks are natural function approximators (by the Universal Approximation Theorem), we wish to learn $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$, which takes real-coordinates as input and outputs the RGB value for the pixel. We aid the learning of this function with encodings from SinGAN generators.

4 Methodology

SinGAN training proceeds as is described in the original paper. For training the MLP, we downscale the input image to various scales at every level of the generator pyramid and set up a base coordinate system on the lowest level image. High resolution images from this set are then mapped to the base image system to obtain real-valued pixel coordinates, which forms the training set for the MLP. Also, while training the MLP, we append the input coordinates to the hidden layers as well to ensure that the effect of the coordinates is not lost over the depth of the network.

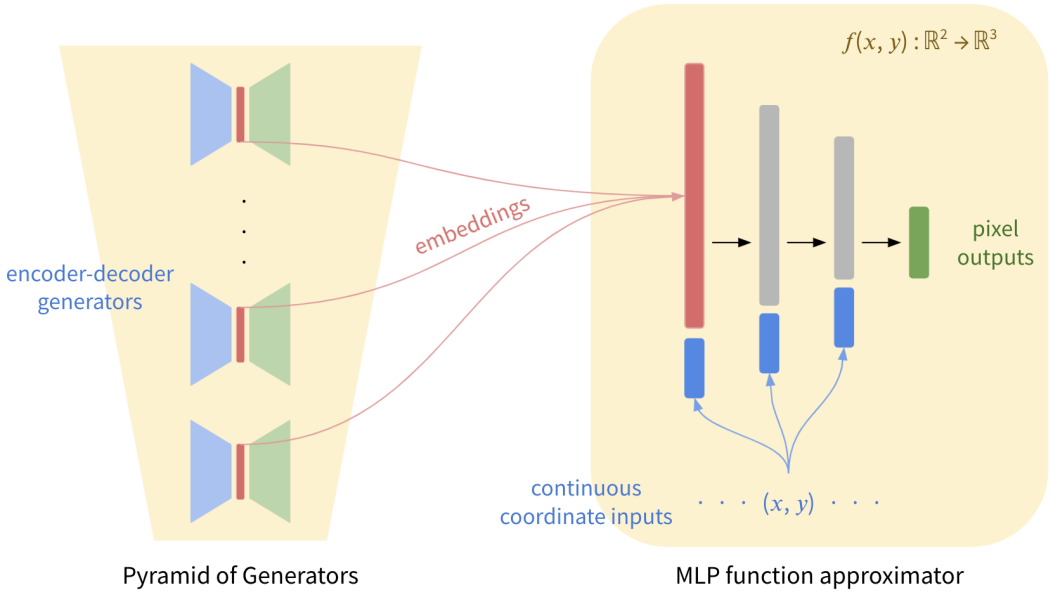
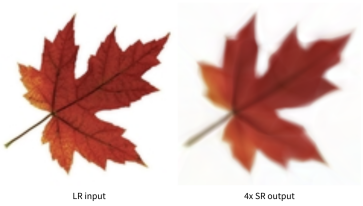


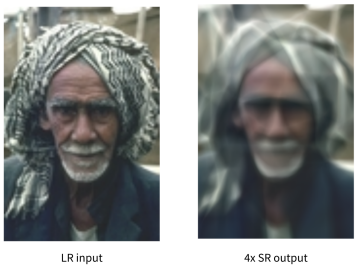
Figure 3: MLP trained with encodings from SinGAN to predict pixel outputs for real-valued coordinate inputs

We train SinGAN for 2000 epochs on MSE loss for both generator and discriminator. Once the pyramid of generators is trained, we extract encodings from the autoencoders at various scales and train the MLP for 1000 epochs on MSE loss. The SinGAN generator and discriminator are trained with a learning rate of 0.0005 with MultiStepLR scheduler and the MLP is training with a learning rate of 0.0001 with StepLR scheduler.

5 Results



(a) 4x super-resolution on simple image



(b) 4x super-resolution on complex image

(c) 4x and 64x super-resolution on complex image

Figure 4: SinGAN + MLP superresolution results

We observe that the overall image-distribution is learnt quite well by the MLP. However, when it comes to the finer details, we observe an undesired smoothing effect (particularly on the edges) persistent across all images. We believe this is due to the fact that we are learning a smooth, continuous function over the co-ordinate space which makes it difficult to make sharp jumps on the edges. The smoothing effect also results in loss of finer details such as the eyes of the man in figure 4(b). We also note that our predictions are consistent over multiple scales of superresolution - 4x and 64x and seen in the figure 4(c) above. Although we see good overall regeneration at higher resolutions, the results are not comparable to the SOTA techniques in superresolution. More results and the generated images can be seen here.

6 Applications

With the trained function approximator, we can query arbitrary pixels at high-resolution. For Edge Selective Super-Resolution, we obtain the list of edge co-ordinates using the Canny-edge detection algorithm and super-resolve only along the edges to try and obtain a higher quality as desired only along the edges.

Figure 5: Edge-selective Super Resolution using SinGAN + MLP function approximator

Due to low-quality results generated, we were unable to demonstrate good results on the original task of 'Edge-Selective Super-Resolution' for a larger set of images. Once the results are sufficiently improved, we can query the image only for the edge pixels to perform super-resolution only along the edges. Another application would be to query the function approximator on foreground image segments to produce an iPhone-style portrait mode effect, where the subject is rendered in high-resolution but the foreground is low resolution (or a simple upscaled version).

7 Conclusions

We summarize that the current state of the model is good enough to learn higher-level features but misses out on the fine-grained details. The function approximator approach does show promising results. As a future work, we can make the results better by improving the quality of encodings using deeper autoencoders, adding skip-connections (as a U-Net) in the SinGAN generators. We can also try a deeper MLP for function approximation.

The source code for the implementation can be found here.

References

Rouf, M. et al. (2016). "Fast edge-directed single-image super-resolution". In: *Image Processing: Algorithms and Systems*.

Nazeri, Kamyar, Harrish Thasarathan, and Mehran Ebrahimi (2019). "Edge-Informed Single Image Super-Resolution". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3275–3284.

Yu, X. and F. Porikli (2016). "Ultra-Resolving Face Images by Discriminative Generative Networks". In: *ECCV*.

Wong, Lone et al. (2020). "Perceptual Image Super-Resolution with Progressive Adversarial Network". In: *ArXiv* abs/2003.03756.

Chen, Yu et al. (2018). "FSRNet: End-to-End Learning Face Super-Resolution With Facial Priors". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2492–2501.

Keys, R. (1981). "Cubic convolution interpolation for digital image processing". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* vol. 29, no. 6, pp. 1153–1160.

Duchon, C. E. (1979). "Lanczos filtering in one and two dimensions". In: *Journal of Applied Meteorology* vol. 18, no. 8, pp. 1016–1022.

Yang, Chih-Yuan, Chao Ma, and Ming-Hsuan Yang (2014). "Single-Image Super-Resolution: A Benchmark". In: *ECCV*.

Kim, Jiwon, J. Lee, and Kyoung Mu Lee (2016). "Accurate Image Super-Resolution Using Very Deep Convolutional Networks". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654.

Sajjadi, Mehdi S. M., B. Schölkopf, and M. Hirsch (2017). "EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4501–4510.

Soh, Jae Woong et al. (June 2019). "Natural and Realistic Single Image Super-Resolution With Explicit Natural Manifold Discrimination". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shaham, Tamar Rott, Tali Dekel, and T. Michaeli (2019). "SinGAN: Learning a Generative Model From a Single Natural Image". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4569–4579.

Jiang, Chiyu "Max" et al. (2020). "MESHFREEFLOWNET: A Physics-Constrained Deep Continuous Space-Time Super-Resolution Framework". In: *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–15.