A Study of Perceptually Effective Image Super Resolution Models

Zihan Chen, Tian Jian Wang, Hongbo Zhang, Xiaoji Zhang
University of Waterloo
{zihan.chen, tj3wang, hongbo.zhang, joy.zhang.1}@uwaterloo.ca

1 Introduction

Single-image super resolution (SISR) is a classical problem in computer vision that aims at recovering a high-resolution image from its low-resolution counterpart, which offers the promise of overcoming the inherent resolution limitations of low-cost imaging sensors, allows better utilization of high-resolution displays, and is essential in medical imaging and satellite imaging where diagnosis and analysis require high-resolution images [1].

In this report, we focus on the super resolution problem for human perception. Specifically, we want to find a computationally efficient technique that generates high-resolution images of perceptually high accuracy. We evaluate the generated high-resolution images using the Peak Signal-to-Noise Ratio (PSNR), which computes the pixel-wise independent mean-squared error between an output image and the ground truth. Additionally, we are interested in whether the generated images are perceptually satisfying, i.e. contain high-frequency details that match the fidelity expected at the higher resolution [8].

We start by introducing the previous work in image super resolution. Subsequently, we compare four SISR methods (interpolation, sparse coding, convolutional neural network, generative adversarial network) in terms of training efficiency and quality of the output images. We then conduct experiments to further evaluate the performance of the SISR models both quantitatively and perceptually. Finally, we conclude that the convolutional neural network model produces the best computational efficiency, and the generative adversarial network model generates images of the best perceptual quality.

2 Related Work

The super resolution image reconstruction is a severely ill-posed problem in that there exists a multiplicity of solutions given the insufficient number of low-resolution images and the unknown blurring operators [2]. To stabilize the inversion of the ill-posed problem, numerous regularization techniques have been proposed [3]. However, the performance of regularization is degraded with the decreasing number of input images, and the output images lack high-frequency details [2].

Another class of approaches is based on interpolation, where the unknown pixels in an upscaled image are filled with the weighted averages of their nearest neighbours. While simple interpolation methods (e.g. biliear and bicubic) generate overly smooth images, interpolations that exploit the prior of natural images (e.g. explore the gradient profile of local patches [4], distinguish between foreground/background [5]) are capable of preserving the edges in the zoomed image.

Several recent methods adopt the example-based strategy to constrain the solution space of the SISR problem with strong prior information [6]. The most representative example-based approach is the sparse-coding based method, which encodes overlapping patches of the input image by a low-resolution dictionary, passes the sparse coefficients into a jointly-trained high-resolution dictionary for reconstruction, and aggregates the reconstructed high-resolution patches to produce the output [1]. The approach is shared among multiple example-based methods, where optimization is performed on

choosing the optimal dictionary size and patch size. However, other steps in the approach have not been considered in a unified framework [6].

In addition to the example-based strategies, convolutional neural network (CNN) based SISR models have shown stellar performance in terms of efficiency and accuracy. Wang et al. demonstrated that the sparse-coding SISR model can be incarnated as a deep neural network, which leads to more effective training [7]. Dong et al. used bicubic interpolation to upscale an input image and trained a 3-layer deep convolutional network end-to-end to obtain the mapping between the low and high resolution images, which achieved excellent performance compared to the aforementioned methods. [6]

To recover the finer texture details when super resolving at large upscaling factors, Ledig et al. presented a generative adversarial network (GAN) for SISR, where the generator consists of a deep residual network that can recover photo-realistic features from heavily downsampled images, and the discriminator is trained to differentiate between the super resolved images and the original photo-realistic images [8]. Furthermore, Ledig et al. proposed a perceptual loss function to evaluate perceptual similarity (instead of pixel space similarity, as in PSNR) between images. SRGAN has been proved to achieve state-of-the-art performance for large upscaling factors, and the images it generates are more photo-realistic than those produced using previous methods, according to the perceptual loss benchmark [8].

3 Methods

In this section, we introduce four SISR techniques, namely, interpolation, sparse representation, Super Resolution Convolutional Neural Network (SRCNN), and Super Resolution Generative Adversarial Network (SRGAN). We compare the effectiveness of the models in terms of training efficiency and quality of the super-resolved images.

3.1 Interpolation

Interpolation is a standard super resolution technique, which is widely used given its simplicity and relatively satisfying performance. Traditional polynomial-based interpolation methods include bilinear and bicubic interpolations, where an input low resolution image is upscaled, and each unknown pixel in the zoomed image is filled with the weighted average of its neighbouring pixels (bilinear interpolation averages the 4 nearest neighbours and bicubic interpolation uses the 16 nearest ones). While interpolation methods preserve the content of the original image and are easy to implement, they tend to generate overly smooth images with blurring and ringing artifacts [2], which is a major disadvantage considering our goal of generating images of high frequency details.

In addition to the pure interpolation, we also consider a combination of dimensionality reduction and interpolation methods to cope with the presence of noisy images. In particular, we perform Principle Component Analysis (PCA) and Non-negative Matrix Factorization (NMF) on the low-resolution images before applying bicubic interpolation. Despite the contribution to noise reduction, the dimensionality reduction methods are not particularly good at super-resolving images, in that they drop information in the low-resolution images when performing matrix factorization.

Despite the perceptually unsatisfying images it generates, interpolation is still used as a baseline in many works regarding image super resolution, given its ubiquity in the field and the high PSNR value it produces (as shown in later experiments).

3.2 Sparse Representation

Given that the interpolation-based methods generate overly smooth images, we wish to find a model that generates images with more detailed features and clearer boundaries. Several recent methods achieve this goal by adopting the example-based strategy, which utilizes a database consisting of high- and low-resolution image patch pairs, and aims at either exploiting the similarities of the same image or learning a mapping from the patch pairs in the training set [6].

The most representative example-based SR method is the sparse coding based method, with the idea of representing the image patches as a sparse linear combination of elements from an over-complete dictionary. By enforcing the sparse representation similarity between the low- and high-resolution image patches via jointly training two dictionaries for the low- and high-resolution image patches

respectively, we can apply the sparse representation for each low-resolution image patch with the high-resolution patch dictionary to generate the corresponding high-resolution patches, and aggregate the patches to get the resulting super-resolved image [2].

To apply sparse coding to the SISR problem, there are two constraints: the reconstruction constraint, i.e. the input low-resolution image is a downsampled version of the original high-resolution image, and the sparsity prior constraint, meaning that each patch in the high-resolution image has a sparse representation from the high-resolution image patch dictionary [2]. Furthermore, local consistency is guaranteed by allowing overlapping patches in the input low-resolution image and requiring that the output high-resolution patches agree on the overlapping areas [2].

The sparse coding algorithm starts with extracting each 3×3 patch y in the input low-resolution image $\mathbf{Y}$ with 1 pixel overlap between adjacent patches. Subsequently, an optimization problem is solved with respect to $\mathbf{D}_h$ and $\mathbf{D}_l$ (corresponding to the high- and low-resolution image patch dictionaries) to find the sparse representation α of y in $\mathbf{D}_l$. The corresponding high-resolution patch x is then generated by applying α with $\mathbf{D}_h$. Upon aggregating all the generated high-resolution patches and producing the high-resolution image $\mathbf{X}_0$, gradient descent is used to find the closest image $\mathbf{X}$ to $\mathbf{X}_0$ that satisfies the reconstruction constraint, and $\mathbf{X}$ is the final output of the algorithm [2].

Unlike other example-based SR methods that directly sample from the large number of high- and low-resolution image patch pairs, the sparse coding method learns a compact representation (i.e. the dictionaries) from the patch pairs to compute the co-occurrence prior [2], which greatly improves the speed of the algorithm. Despite the ill-posed condition of the SISR problem, the sparse coding method is able to regularize the problem efficiently and robustly.

3.3 Super-Resolution Convolutional Neural Network

The convolutional neural network model for super-resolution (SRCNN) follows closely from the aforementioned sparse coding based method, and directly learns an end-to-end mapping between the low- and high-resolution images [6]. In spite of the motivation from the example-based methods, the convolutional model does not learn the dictionaries for the image patches, but instead uses hidden layers in the neural network to model the patch space and conduct patch extraction/reconstruction at the start/end of the solution pipeline, followed from the sparse coding model [6].

The structure of the SRCNN model is elaborated as follows (also shown in Figure 1, which is adapted from the original paper). Given an input low-resolution image, the first layer extracts patches from the image and generates a high-dimensional vector representation for each patch, which comprise a set of feature maps [6]. The second layer then non-linearly maps each vector in the feature maps to a high-resolution patch representation. Finally, the last layer reconstructs the high-resolution image by combining the predictions in a spatial neighbourhood [6]. The three layers are combined to form a CNN. Additionally, the model uses the Mean Squared Error (MSE) as the loss function, which is partially related to the perceptual quality of the output images. It is worth mentioning that the SRCNN model can flexibly adapt to other evaluation metrics as long as they are derivable, which is hard to achieve using the aforementioned methods.

The SRCNN solution pipeline is extremely similar to that of the sparse-coding based method. However, the sparse-coding method only concerns the optimization of the high- and low-resolution image patch dictionaries, whereas the SRCNN model optimizes an end-to-end mapping consisting of all operations, which includes not only the dictionaries, but also the non-linear mapping, the mean subtraction and the averaging step. Despite its lightweight structure, the SRCNN model achieves superior performance than the sparse coding method in terms of training speed and the sharpness of the super-resolved images.

3.4 Super-Resolution Generative Adversarial Network

Despite the vast improvement in speed and accuracy of SISR using deep convolutional networks, it is still unknown how we can recover the fine texture details when super-resolving at large upscaling factors [8]. Instead of using optimization-based methods with the objective of maximizing the PSNR value (i.e. minimizing the mean-squared reconstruction error), a more natural choice is to apply the notion of generative modeling, which requires the model to insert more information into the original input image, as in the goal of the SISR problem.

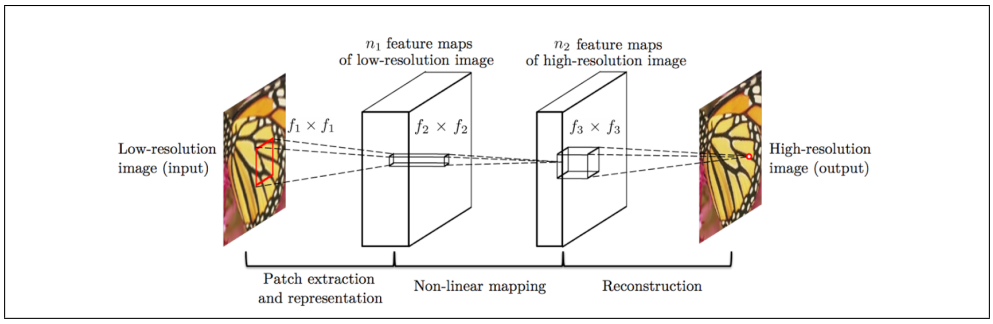

Figure 1: An illustration of the SRCNN model. *Adapted from "Image super-resolution using deep convolutional networks" by Dong, Chao et al., 2016, IEEE transactions onpattern analysis and machine intelligence38.2: 295-307. Copyright 2016 by the Name of Dong, Chao et al.. Adapted with permission.*

A new framework for estimating generative models is the generative adversarial network (GAN), which is powerful in generating images with high perceptual quality that match the expected fidelity, in that it encourages the solutions to move towards the natural image manifold that contains photo-realistic images with high probability [8]. In particular, the GAN model for super-resolution (SRGAN) defines an adversarial min-max problem between a generative network and a discriminative network with the goal of optimizing a perceptual loss function, where the generator G is trained with the goal of fooling the discriminator D that is trained to distinguish between the super-resolved reconstructed images and the original high-resolution images. Furthermore, the perceptual loss function assesses the super-resolved image based on perceptually relevant characteristics [8] that are not captured by metrics such as MSE and PSNR.

The structure of the SRGAN model is as follows. The generator is constructed using a deep residual network with 15 residual blocks, each block consisting of 2 covolutional layers, 64 feature maps, batch normalization layers and ParametricReLu as the activation function [8]. Subsequently, two subpixel convolutional layers are trained to increase the input image resolution. By applying residual learning, each layer can fine-tune the output of the previous layer by adding a "residual" to the input, which drives each new layer to learn additional details of the image in the case of SISR. The discriminator consists of 8 convolutional layers, uses strided convolutions to reduce the image resolution every time the number of features is doubled, and applies two dense layers as well as a sigmoid activation function to produce the probability for classification [8].

Unlike the aforementioned methods that utilize the standard quantitative measures (e.g. MSE and PSNR) for evaluation and focus on the computational efficiency of the SR algorithms, SRGAN emphasizes the perceptual quality of the super-resolved images by introducing a perceptual loss function that captures the visual characteristics of the output images, thus generates photo-realistic images and acheives state-of-the-art for SISR with high-scaling factors (as measured by PSNR and structural similarity) [8].

4 Experiments

In this section, we demonstrate our experimenting process with respect to the aforementioned methods, and analyze the model performance as well as the quality of the super-resolved images both quantitatively and perceptually.

4.1 Data

We use a subset (1000 images) of the Large-scale CelebFaces Attributes (CelebA) data set [9] for our experiment. We start by downsampling the original images to obtain the corresponding low-resolution images, and then we upscale the low-resolution images by a factor of 3 during the reconstruction, which is commonplace in the literature concerning the SISR problem [2].

Figure 2: Results of the woman image magnified by a factor of 3 and the corresponding PSNR values. Left to right: input, Bilinear interpolation (PSNR:25.47), Bicubic interpolation (PSNR:26.19)

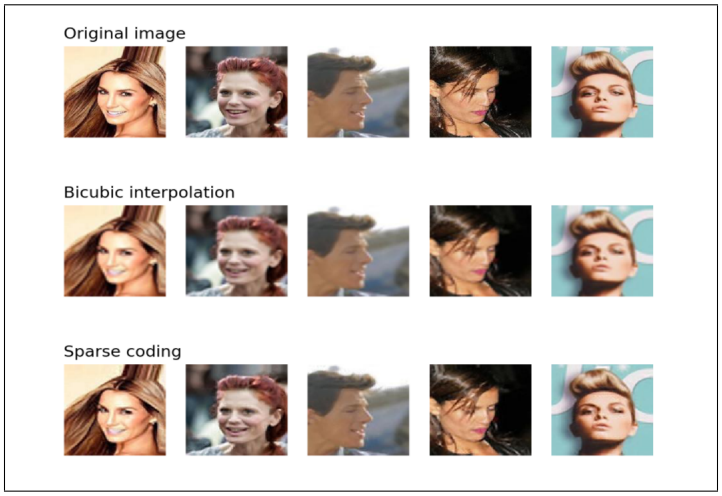

Figure 3: A comparison of Bicubic interpolation and sparse-coding based image SR.

4.2 Result and Analysis

The results of the interpolation methods are shown in Figure 2. Despite the high PSNR score, the super-resolved images appear overly smooth and lack high-frequency details, which agrees with our previous argument regarding the limitations of interpolation.

The sparse-coding based SR model generates sharp results with high-frequency details, as shown in Figure 3. However, the training process is extremely slow, given that the optimization procedures for patch detection and sparse coding are computationally intense.

The convolutional neural network model also generates super-resolved images that preserve the original information and produces sharp edges without any obvious artifacts across the image [6] (Figure 4). Additionally, it achieves a higher PSNR score than interpolation, and is trained much faster than the sparse-coding based method. Compared to other CNN structures, a relatively lightweight structure is adopted to achieve this result, and we believe that the perceptual quality of the output images can be improved on a larger network scale.

The training of the SRGAN model is also slow, given its complex structures and the large number of parameters it optimizes. However, Figure 5 shows significant improvement in the image quality over the first 20 epochs. Furthermore, the model achieves a higher PSNR score than interpolation and sparse representation on some images, providing us with enough evidence to believe that the SRGAN method can produce a more convincing result with a larger number of epochs.

As shown in Table 1, in spite of the more perceptually accurate images they generate, the state-of-the-art SR methods (the latter 3 in the table) do not produce better PSNR values than the standard Bicubic interpolation method, which is only able to generate overly-smooth images of poor perceptual quality. The PSNR value is therefore ineffective in capturing the perceptual characteristics of the super-

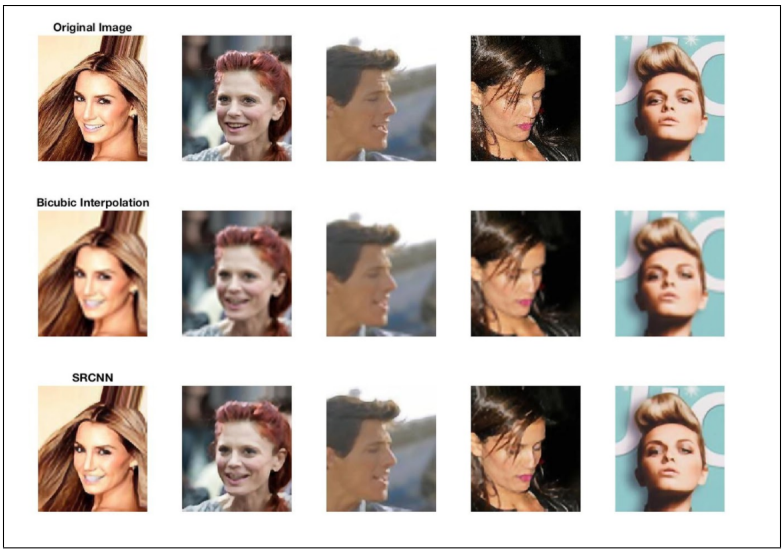

Figure 4: Comparison of Bicubic interpolation and convolutional neural network for image SR.

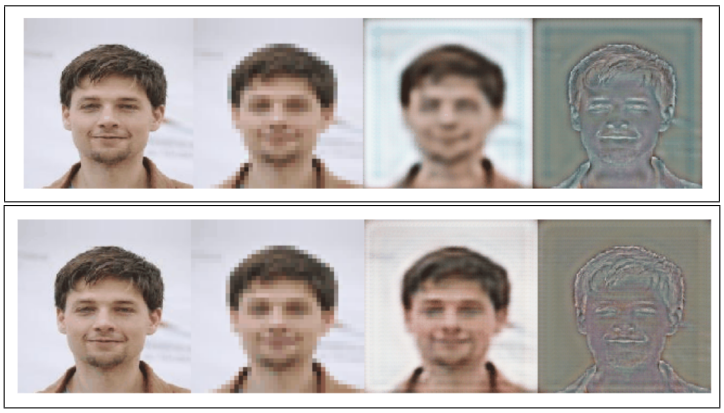

Figure 5: Left to right: original image, input, SRGAN and the difference between the SRGAN super-resolved image and the original. The first row is the result after the first 5 epochs, and the second row is the result after 20 epochs.

resolved images, and additional benchmarks should be used to evaluate the perceptual effectiveness of the SISR models.

5 Conclusion

We have introduced four methods of single-image super resolution, namely, interpolation, sparse coding, convolutional neural network (SRCNN) and generative adversarial network (SRGAN). We have compared the methods in terms of training efficiency and the quality of the images generated,

PSNR	image1	image2	image3	image4	image5
Bicubic	23.87	28.48	36.24	24.56	29.98
Sparse Representation	26.28	28.09	33.62	28.93	28.79
SRCNN	24.97	29.70	37.58	25.22	31.60
SRGAN	23.92	28.15	32.52	23.21	27.83

Table 1: PSNR score of all methods on five images

both quantitatively using the PSNR metric and perceptually by observation. We have shown that SRCNN is the most computationally efficient method, and SRGAN is the most perceptually effective method, in terms of the restoration of high-frequency details in the reconstructed images and the expected fidelity at the high-resolution counterpart of the input images. Furthermore, we have claimed that the PSNR values have no correspondence with the perceptual quality of the super-resolved images. Future effort is required in terms of training the SRGAN model for a larger number of epochs to produce more convincing and comparable results against other methods, as well as proposing a new benchmark for evaluating the SR models that accounts for the perceptual characteristics of the generated images.

References

[1] Yang, Jianchao, et al. "Image super-resolution as sparse representation of raw image patches." *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 2008.

[2] Yang, Jianchao, et al. "Image super-resolution via sparse representation." *IEEE transactions on image processing* 19.11 (2010): 2861-2873.

[3] M. E. Tipping and C. M. Bishop, "Bayesian image super-resolution," in Advances in Neural Information and Processing Systems 16 (NIPS), 2003.

[4] Sun, Jian, Zongben Xu, and Heung-Yeung Shum. "Image super-resolution using gradient profile prior." *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 2008.

[5] Dai, Shengyang, et al. "Soft edge smoothness prior for alpha channel super resolution." *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on.* IEEE, 2007.

[6] Dong, Chao, et al. "Image super-resolution using deep convolutional networks." *IEEE transactions on pattern analysis and machine intelligence* 38.2 (2016): 295-307.

[7] Wang, Zhaowen, et al. "Deep networks for image super-resolution with sparse prior." *Proceedings of the IEEE International Conference on Computer Vision.* 2015.

[8] Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." *arXiv preprint arXiv:1609.04802* (2016).

[9] "Large-Scale Celebfaces Attributes (Celeba) Dataset". Mmlab.ie.cuhk.edu.hk. N.p., 2017. Web. 22 Apr. 2017.