SR-BiGAN: Exploring Image Super-resolution with Generative Model

Chengliang Tang
Department of Statistics
Columbia University
New York, NY 10027
ct2747@columbia.edu

May 18, 2020

Abstract

Image super-resolution is of great importance in a variety of digital imaging applications, and numerous algorithms using deeper convolutional neural networks have been developed in the last decade. Yet most of the state-of-the-art methods depend on an artificially created training set where low-resolution images are down-sampled by interpolation and blurring, and thus fall short of generating high-quality outputs in real-world applications. In this project, we explores learning image super-resolution with generative models. Motivated by the architecture of BiGAN, we propose SR-BiGAN, an end-to-end learning algorithm for image super-resolution with weak supervision. This algorithm simultaneously trains trains low-to-high super-resolver (decoder) and a high-to-low down-sampler (encoder). With preliminary results on remote sensing data (due to time limit), we show this framework is able to generate close-to-real high-resolution images, but suffer from information loss in object labels.

1 Introduction

In many digital imaging applications, high-resolution images are necessary for research analysis, while are usually unavailable due to the technical limitations of imaging devices. For example, recently developed remote sensing techniques such as aerial imagery have enabled efficient gathering of high-quality images over large spatial scale. With those high-resolution data, numerous important characteristics could be extracted from images. Nonetheless, historical data captured decades ago cannot enjoy such advantage due to the low-resolution. As a result, a key question in computer vision is how to generate high-resolution images from low-resolution input.

Image super-resolution is an important computer vision task with many interesting applications, such as movie standard conversion, face recognition, etc. With the recent development of deep learning, image super-resolution algorithms using deeper convolutional neural networks have been proposed in the last decade [2]. Replacing the interpolation processing with convolutional layers, these algorithms are able to learn image super-resolution in an end-to-end, scalable framework with stochastic gradient descent. Using deeper neural networks, most of the recent advances in image super-resolution require a large training set for model fitting. And in practice, training set are created by simply down-sampling the high-resolution images to generate the low-resolution inputs. However, strong prior knowledge are already integrated in the training set creation since the real-world low-resolution images are not necessarily generated in this way. As a result, these algorithms cannot be directly applied to real-world tasks and fell short to generate visually close-to-real high-resolution images. In this paper, we tackle the challenge of image super-resolution with SR-BiGAN, an end-to-end framework to generate high-resolution images. Without generating low-resolution images using simple downsampling, we build an encoder to learn the high-to-low down-sampling process and fit it simultaneously with the super-resolution generator.

2 Motivation

The motivation of this project is driven by the need in my Ph.D. research project. I have two large-scale aerial image datasets of El Yunque National Forest in Puerto Rico (more details introduced in Experiments). One of them (captured by drones) is of high-resolution, but has a relatively small spatial coverage. And the other one (captured by satellites) is of low-resolution, and enjoys a much larger spatial coverage. Therefore, if we could learn a reliable image super-resolver and apply it to the low-resolution-large-coverage dataset, more valuable forest characteristics would be mined for environmental analysis.

From a representation learning perspective, the learning of an image super-resolver is equivalent with the inverse process of inferring the original "data" (high-resolution images) from their "representations" (low-resolution images). Therefore, a number of existing algorithms in generative models, e.g. GANs, VAEs, could be helpful to the research goal.

3 The challenge of weak supervision in environmental studies

In environmental studies, historical data are as important as those recently collected because they provide a comprehensive description of the time-varying environmental factors. Usually, data collected at different timestamps would have different spatial coverage and exhibit multiple visual patterns. Figure 1 illustrates the most common case in remote sensing imagery data analysis. Low-resolution images are available in a large spatial scale (area $A + B$). In contrast, high-resolution images collected with more cost are only obtainable in a different small subset (area $B + C$) with few overlapping (area B). While, image pairs on overlapping area B might not perfectly match since the time gap between their collections vary from years to decades.

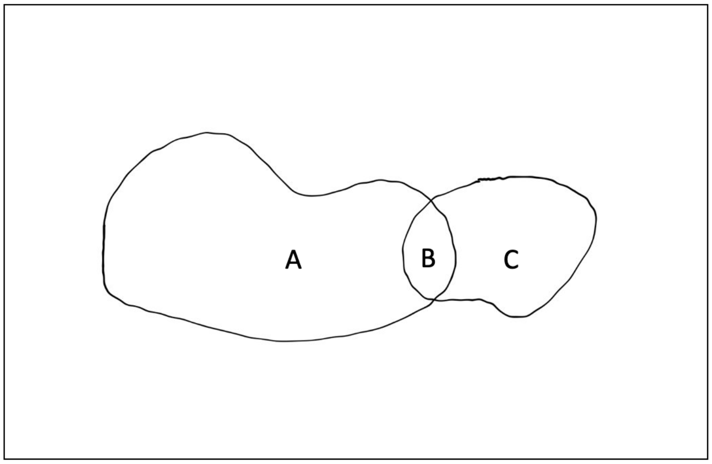

Figure 1: Spatial coverage of data in different resolutions. A: areas with only low-resolution data available. C: areas with only high-resolution data available. B: overlapping areas with both low/high-resolution data available.

This case is also an example of weak supervision, where the training labels are noisy (inaccurate supervision) and partially available (incomplete supervision). In this scenario, simply down-sampling the high-resolution image for model training is not the most ideal case because there are some unknown factors influencing the quality of low-resolution aerial images, e.g. illuminance, reflectance.

4 SR-BiGAN: Image Super-resolution Motivated by BiGAN

As a variant of the GAN model, BiGAN proposed by Donahue et al. provides a means of learning the inverse mapping of the generator. In their framework, the encoder and the generator are trained together to "fool" the discriminator, and the optimal solution can be achieved when they perfectly recover the conditional distributions. In our proposed algorithm, we would adopt the major structure of BiGAN and adapt it to solve the weak supervision challenge.

4.1 Notations and Assumptions

We denote the low-resolution images as I^{LR}, which is a real-valued tensor of size $W \times H \times C$, and the corresponding high-resolution image as I^{HR}. The training data is composed of two sets, which corresponds to the areas with high-resolution images in Figure 1. The first component of training data is $\mathcal{B} = \{I_i^{LR}, I_i^{HR}\}_{i=1}^{N}$, which is composed of N pairs of LR-HR images. The second component is $\mathcal{C} = \{I_j^{HR}\}_{j=1}^{M}$, which is composed of M high-resolution images. And the goal is to make predictions for the test set $\mathcal{A} = \{I_k^{LR}\}_{k=1}^{K}$, which contains K unlabeled low-resolution images. Here, $K \gg M \gg N$.

Suppose $P(I^{HR})$ and $P(I^{LR})$ are the marginal distributions of high-resolution images and low-resolution images. $P(I^{LR}|I^{HR})$ and $P(I^{HR}|I^{LR})$ are the two conditional distributions. In our framework, the key assumption is those distributions are identical in the three datasets $\mathcal{A}, \mathcal{B}, \mathcal{C}$. This assumption might be too strong for the environmental studies, since factors are likely to vary along different areas. But how to incorporate spatial information into the image analysis is beyond the scope of this paper.

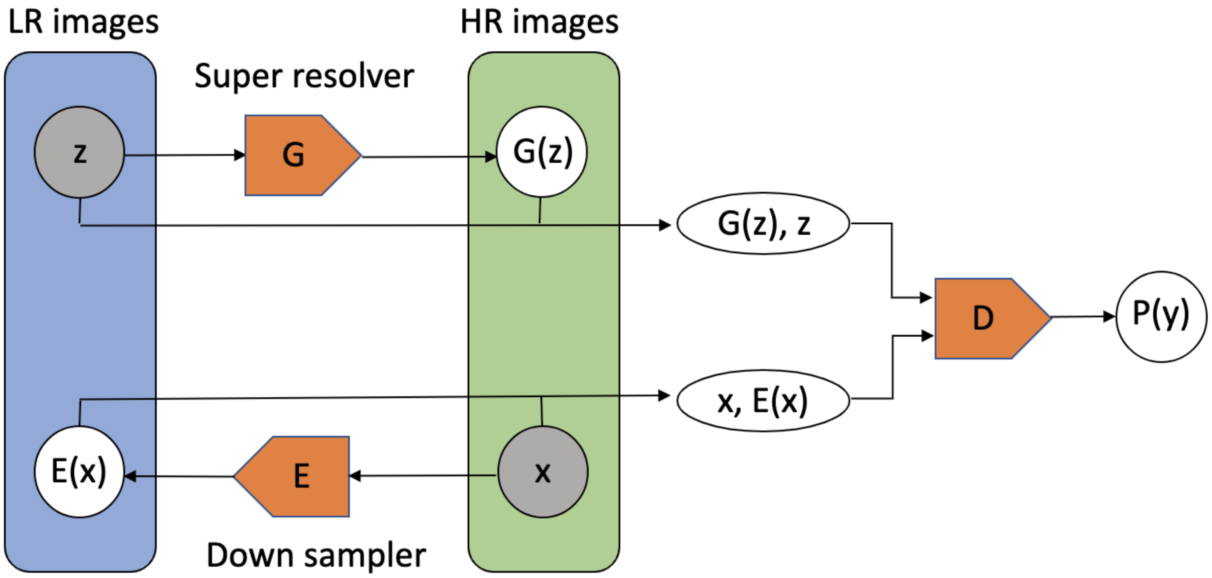

Figure 2: Spatial coverage of data in different resolutions. A: areas with only low-resolution data available. C: areas with only high-resolution data available. B: overlapping areas with both low/high-resolution data available.

4.2 Algorithm

In the vanilla GAN model, fake images are generated by applying a differentiable neural network over random vectors. A natural idea is to use the low-resolution images as the "seed" for the generator, and therefore the framework of GAN can be directly transferred to the task of image super-resolution. Such idea is broadly adopted in many computer vision applications of GAN, such as SRGAN.

Figure 2 illustrates the overall model of SR-BiGAN. In addition to the generator (super resolver) G from the standard SRGAN framework, SR-BiGAN includes and encoder (down sampler) E which maps the high-resolution images $x = I^{HR}$ back to the low-resolution $z = I^{LR}$. And the discriminator D is defined over the joint space of I^{HR} and I^{LR}. This framework is motivated by the structure of BiGAN, and the training objective is defined as a minimax objective

$$\min_{G,E} \max_{D} V(D, E, G) \tag{1}$$

where

$$V(D, E, G) = \mathbb{E}_{\mathbf{I^{HR}} \sim P_{\mathbf{HR}}} \big[\log D(\mathbf{I^{HR}}, E(\mathbf{I^{HR}})) \big] + \mathbb{E}_{\mathbf{I^{LR}} \sim P_{\mathbf{I_{LR}}}} \big[\log(1 - D(G(\mathbf{I^{LR}}), \mathbf{I^{LR}})) \big] \tag{2}$$

Theoretical results in BiGAN shows the optimal encoder E and the generator G should learn to invert each other. In the optimal case, they perfectly recover the conditional distribution $P(I^{LR}|I^{HR})$ and $P(I^{HR}|I^{LR})$.

Sample confidence To our knowledge, the idea of learning the down-sampler from training images was first proposed by Bulat et al. In their framework, they first fit a high-to-low SRGAN model using the unmatched low-high image pairs, use the SRGAN to generate low-resolution images, and finally learn another low-to-high SRGAN model using the generated image pairs. However, since the SRGAN model can never perfectly learn the marginal distribution of low-resolution images. The generator is unable to communicate with the real low-resolution images in that case, which would limit its generalization ability.

Therefore, in our framework, sample reliability was introduced to decide whether use the real low-resolution image or the generated low-resolution image. The idea is quite straightforward: for every sample image pair (I_i^{LR}, I_i^{HR}) in our training component $\mathcal{B} = \{(I_i^{LR}, I_i^{HR})\}_{i=1}^{N}$, we would assign them a confidence score r_i, which is measures their degree of match and is calculated as the *correlation coefficient* between their pixel-values. For those image pairs with a high confidence score, we would input the real low-resolution image for training the generator G and the discriminator D. And for those images with a low confidence score, we would use the fake low-resolution images $E(I^{HR})$ for training the generator G and the discriminator D.

In practice, a threshold $\hat{r}$ is chosen for the confidence score, and thus the training set is divided into three components: $\tilde{\mathcal{B}}$ of high-confident image pairs, $\mathcal{B} \setminus \tilde{\mathcal{B}}$ of low-confident image pairs, and $\mathcal{C}$ of high-resolution images. In short, the generator and discriminator are trained on the whole training dataset $\mathcal{B} \cup \mathcal{C}$, but the encoder is only trained on $\tilde{\mathcal{B}}$.

Perceptual loss PSNR is the most commonly used loss function in image super-resolution. While, they are unable to measure the semantic similarity between multiple images. Perceptual loss proposed by Ledig et al. is a weighted sum of the content loss and the adversarial loss, and show the advantage of generating sharp and visually pleasing images. To be specific, the content loss is calculated as the MSE in the feature space of fully-connected layers of VGG19, and minimizing the adversarial loss drives the generator to fool the discriminator.

In our SR-BiGAN framework, the generator G is optimized over the perceptual loss, and the encoder E is optimized only using the adversarial loss.

Training details The training of GAN is notoriously unstable. We should swap the learning objectives and alternate between updating generator and discriminator to provide a strong gradient.

Algorithm 1 SR-BiGAN with weak supervision

```
Parameters:
$\hat{r}$: the threshold of sample confidence
num_iter: maximal number of iterations
Input:
Training data: $\mathcal{B} = \{I_i^{LR}, I_i^{HR}\}_{i=1}^N$ and $\mathcal{C} = \{I_j^{HR}\}_{j=1}^M$
Start training:
Step 1: using threshold $\hat{r}$, split training set into $\tilde{\mathcal{B}}, \mathcal{B} \setminus \tilde{\mathcal{B}}, \mathcal{C}$.
Step 2: initialize generator $G$, encoder $E$ and discriminator $D$.
Step 3: adversarial training
for iter in range(num_iter) do
    Step 4: load training batch $s_1 \subset \tilde{\mathcal{B}}$ and $s_2 \subset (\mathcal{B} \setminus \tilde{\mathcal{B}}) \cup \mathcal{C}$.
    Step 5: generate low-resolution images for $s_2$ using encoder $E$.
    Step 6: update generator $G$ using $s_1 \cup s_2$.
    Step 7: update encoder $E$ using $s_1$.
    Step 8: update discriminator $D$ using $s_1 \cup s_2$ every 30 steps.
End
Output: generator $G$, encoder $E$ and discriminator $D$
```

In SR-BiGAN, there is the same issue. In practice, we usually update the generator G and the encoder E more often than updating D to ensure a relatively weak discriminator.

The training details can be found in Algorithm 1.

5 Experiments

In this part, we apply the proposed SR-BiGAN model to the remote sensing aerial images from Puerto Rico, and the task is to create super-resolved images at 4x scale.

5.1 Datasets

In March 2017, the conditions of El Yunque rainforest in Puerto Rico were characterized at the landscape scale by high-resolution remote sensing data using NASA's G-LiHT platform. G-LiHT is a unique airborne instrument package that acquires **high-resolution** air photo data (pixel size: 3 cm × 3 cm). Besides, **low-resolution** images (pixel size: 10 cm × 10 cm) were collected in Year 2000 with a larger spatial coverage.

In this project, data set is constructed using the 3cm x 3cm high-resolution images. We randomly sample 15,000 high-resolution images (size: 100 x 100), and create the low-resolution training images (size: 25 x 25) by bi-cubic interpolation followed with Gaussian blur. Afterwards, 5,000 images are taken for testsing, and 10,000 images are taken for training with $|\mathcal{B}| = 5,000$, $|\mathcal{C}| = 5,000$.

5.2 Models

In the experiment, we trained our models on a standard NVIDIA K80 GPU. With limited memory, we did not use very deep neural networks for each model. The generator G in the experiment is composed of 4 residual blocks followed by one up-sampling layer. The encoder E is made up of 4 convolutional layers with batch normalization. And for the discriminator D is created with 8 convolutional layers with Leaky ReLU activation, because we believe its discrimination power controls the upper limit of the generator and the encoder. And during the training, we choose to update the generator G and the encoder E in every step, while update D every 30 steps, which turns out to stabilize the training process.

5.3 Performance Evaluation

In the first part, we evaluate the model performance by PSNR, and compare the result with SRCNN using the same generator. After training each model for 100 epochs, we calculate the PSNR for SR-BiGAN and SRCNN. The results show the test PSNR of SR-BiGAN is **16.23**, but for SRCNN is **24.66**, much larger than SR-BiGAN. This performance difference is probably caused by the loss function used in SR-BiGAN, since SRCNN directly optimize over PSNR but SR-BiGAN optimize a perceptual loss.

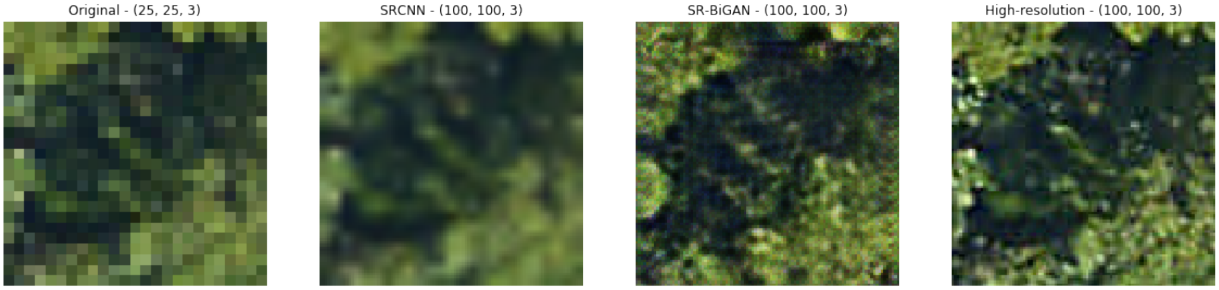

Figure 3: Super-resolved images. First: low-resolution input; Second: SRCNN output; Third: SR-BiGAN output; Fourth: ground truth.

From Figure 3, SR-BiGAN generated more local details than SRCNN, despite the PSNR is much lower.

However, the details recovered by SR-BiGAN are not necessarily true. For example, in Figure 4, part of a palm tree is contained in the image. While, in the generated high-resolution images of SR-BiGAN, the texture of palm trees are replaced by another tree. We believe this phenomenon is caused by “mode collapse” of GAN, and the mode for palm trees is masked in the GAN generator.

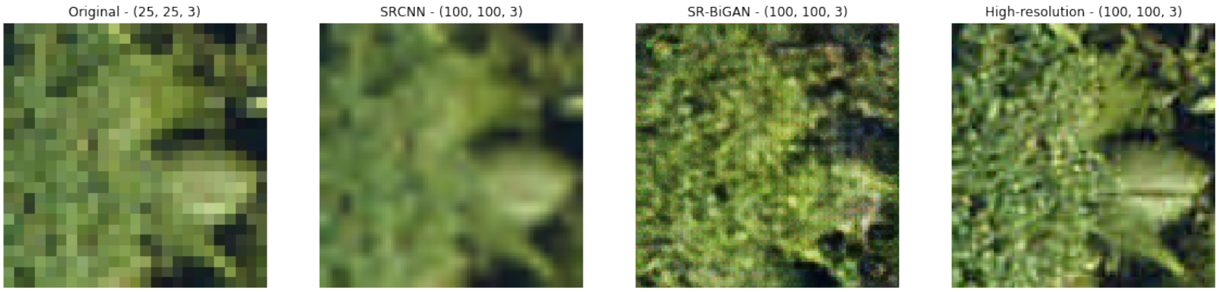

Figure 4: Super-resolved images. First: low-resolution input; Second: SRCNN output; Third: SR-BiGAN output; Fourth: ground truth.

6 Conclusion

In this project, we propose SR-BiGAN, an algorithm for learning image super-resolution with weak supervision. This framework exhibits great flexibility to incorporate information from weak training labels. However, due to time limitation, the results are not good enough, especially the problem of messing with the object labels. Also, in the experiment, we do not have much time to compare it with other weakly supervised image super-resolution algorithms, and the trained model suffer from mode collapse. In future, we would focus more on these two parts, and improve on this framework.

References

[1] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 185–200, 2018.

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[4] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):1127–1133, 2010.

[5] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[6] Karl S Ni and Truong Q Nguyen. Image superresolution using support vector regression. *IEEE Transactions on Image Processing*, 16(6):1596–1610, 2007.

[7] Samuel Schulter, Christian Leistner, and Horst Bischof. Fast and accurate image upscaling with super-resolution forests. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3799, 2015.

[8] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision*, pages 372–386. Springer, 2014.