Pixel x4: Image Super-resolution

Gokul Prasath
Computer Engineering
University of California, Santa Barbara
gokulprasath@ucsb.edu

Venkat Raman
Computer Engineering
University of California, Santa Barbara
venkatraman@ucsb.edu

Vishal Hosakere
Computer Engineering
University of California, Santa Barbara
vishalhosakere@ucsb.edu

Abstract

Increasing photo or image size without losing the quality is nearly impossible. Traditionally photo editing tools like Photoshop, uses Bicubic Interpolation - which makes images smooth and blurry. Though the image resolution is increased, no new information about the image is added to get the extra quality. If we use regular convolutional neural network with pixel to pixel loss function, even a small error would lead to unrealistic images. Generative Adversarial Networks (GANs) like SRGAN [1] are used to upscale the RGB low resolution images to produce visually realistic images than minimizing the discrepancy with the real images. In addition to this, Pixel X4 uses thermal Images along with the low resolution RGB images to enhance the super-resolved image.

Keywords: Generative Adversarial Networks, Perceptual loss, Leaky RELU, Batch normalization

1. Introduction

The recovery of a high resolution (HR) image from its low resolution (LR) counterpart is topic of great interest in digital image processing. This is referred to as super-resolution (SR), finds direct applications in many areas such as HDTV, medical imaging, satellite imaging, face recognition and surveillance. The global SR problem assumes that the LR image is a blurred, down-sampled and noisy version of HR image. It is not a well-posed problem, because the non-invertible lowpass filtering and sampling function induces loss in the high frequency domain. Furthermore, the solution to SR operation is non-trivial as LR to HR is effectively one to many mapping. Thus, this problem has many solutions. Most of the methods usually require computationally complex image registration and fusion stages, the accuracy of which directly impacts the quality of the result. Another family of methods are single image super-resolution (SISR) techniques [2], which seek to learn implicit redundancy present in natural data to recover missing HR information from a single LR instance. This usually arises in the form of local spatial correlations for images and additional temporal correlations in videos. In this case, prior information in the form of reconstruction constraints is needed to restrict the solution space of the reconstruction.

The optimization target of supervised SR algorithms is commonly the minimization of the mean squared error (MSE) between the recovered HR image and the ground truth. This is convenient as minimizing MSE also maximizes the peak signal-to-noise ratio (PSNR), which is a common measure used to evaluate and compare SR algorithms [2]. However, the ability of MSE (and PSNR) to capture perceptually relevant differences, such as high texture detail, is very limited as they are defined based on pixel-wise image differences.

2. Literature Review

Recent overview articles on image SR include Nasrollahi and Moeslund [43] or Yang et al. [61]. Here we will focus on single image super-resolution (SISR) and will not further discuss approaches that recover HR images from multiple images [4, 15]. Prediction-based methods were among the first methods to tackle SISR. While these filtering approaches, e.g. linear, bicubic or Lanczos [14] filtering, can be very fast, they oversimplify the SISR problem and usually yield solutions with overly smooth textures. Methods that put particularly focus on edge-preservation have been proposed [1, 39].

Many methods that are based on example-pairs rely on LR training patches for which the corresponding HR counterparts are known. Early work was presented by Freeman et al. [18, 17]. Related approaches to the SR problem originate in compressed sensing [62, 12, 69]. In Glasner et al. [21] the authors exploit patch redundancies across scales within the image to drive the SR. This paradigm of self-similarity is also employed in Huang et al. [31], where self-dictionaries are extended by further allowing for small transformations and shape variations. Gu et al. [25] proposed a convolutional sparse coding

approach that improves consistency by processing the whole image rather than overlapping patches.

To reconstruct realistic texture detail while avoiding edge artifacts, Tai et al. [52] combine an edge-directed SR algorithm based on a gradient profile prior [50] with the benefits of learning-based detail synthesis. Zhang et al. [70] propose a multi-scale dictionary to capture redundancies of similar image patches at different scales. To super-resolve landmark images, Yue et al. [67] retrieve correlating HR images with similar content from the web and propose a structure-aware matching criterion for alignment.

Neighborhood embedding approaches upsample a LR image patch by finding similar LR training patches in a low dimensional manifold and combining their corresponding HR patches for reconstruction [54, 55]. In Kim and Kwon [35] the authors emphasize the tendency of neighborhood approaches to overfit and formulate a more general map of example pairs using kernel ridge regression. The regression problem can also be solved with Gaussian process regression [27], trees [46] or Random Forests [47]. In Dai et al. [6] a multitude of patch-specific regressors is learned and the most appropriate regressors selected during testing. Recently convolutional neural network (CNN) based SR algorithms have shown excellent performance. In Wang et al. [59] the authors encode a sparse representation prior into their feed-forward network architecture based on the learned iterative shrinkage and thresholding algorithm (LISTA) [23]. Dong et al. [9, 10] used bicubic interpolation to upscale an input image and trained a three layer deep fully convolutional network end-to-end to achieve stateof-the-art SR performance. Subsequently, it was shown that enabling the network to learn the upscaling filters directly can further increase performance both in terms of accuracy and speed [11, 48, 57]. With their deeply-recursive convolutional network (DRCN), Kim et al. [34] presented a highly performant architecture that allows for long-range pixel dependencies while keeping the number of model parameters small. Of particular relevance for our paper are the works by Johnson et al. [33] and Bruna et al. [5], who rely on a loss function closer to perceptual similarity to recover visually more convincing HR images.

3. Proposed Idea

The previous work proposes a super-resolution generative adversarial network (SRGAN) which employs a deep residual network (ResNet) with skip-connection. It defines a perceptual loss using high-level feature maps of the VGG network combined with a discriminator that encourages solutions perceptually hard to distinguish from the HR reference images.

In our proposal, the neural network takes a low-resolution image (RGB and its Thermal Counterpart) and predicts the high-resolution image. The idea behind using thermal image is for better feature extraction which can be then used to reconstruct the image better. Thermal images are unaffected by luminous sources and hence will provide good edge detection and hence can make the SR images sharper.

The predicted image can be compared to the real high-resolution image by using pixel distance. The neural network's aim is to minimize the pixel distance between the two. However, using just the pixel distance may not produce visually realistic images. To counter this, GANs - generators and discriminators - are used to produce the output image [3]. To quickly train the deep networks and to avoid the vanishing gradients in a gradient based weight update model, residual network is explored.

4. Method

RGB and thermal LR image pair are given as input to the modal to produce the HR image. Modal consists of a GAN network comprising of generators and discriminators. Our goal is to train a generating function G that estimates for a given LR input image pair its corresponding HR counterpart. To achieve this, we train a generator network as a feed-forward CNN $G_{\theta G}$ parametrized by θ_G. Here $\theta_G = \{W_{1:L}; b_{1:L}\}$ denotes the weights and biases of a L-layer deep network and is obtained by optimizing a SR-specific loss function l^{SR}. For training images I^{HR}_n, n = 1, . . ., N with corresponding I^{LR}_n, n = 1, . . ., N, we solve:

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l^{SR}(G_{\theta_G}(I_n^{LR}), I_n^{HR})$$

In this work we will specifically design a perceptual loss l SR as a weighted combination of several loss components that model distinct desirable characteristics of the recovered SR image.

4.1. Architecture

Generative adversarial network consists of a generator network that produces an upscaled image based on the training data and a discriminator network discriminates the produced image with the original HR image. We define a discriminator network [3] $D_{\theta D}$ which we optimize in an alternating manner along with $G_{\theta G}$ to solve the adversarial min-max problem.

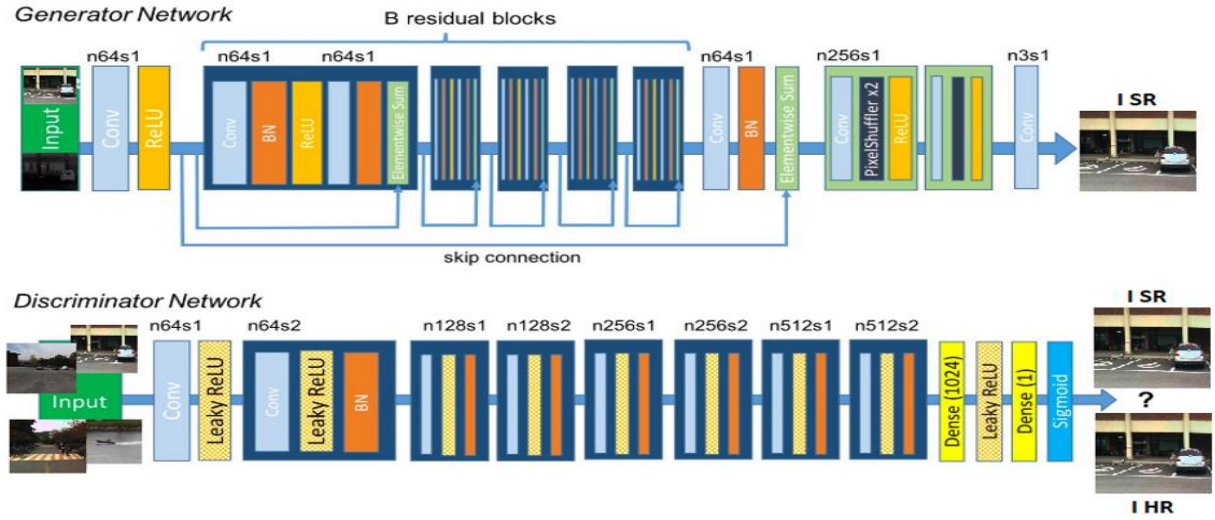

Figure 1: Architecture of Generator and Discriminator Network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I^{HR} \sim p_{\text{train}}(I^{HR})}[\log D_{\theta_D}(I^{HR})] + \mathbb{E}_{I^{LR} \sim p_G(I^{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))]$$

The general idea behind this formulation is that it allows one to train a generative model G with the goal of fooling a differentiable discriminator D that is trained to distinguish super-resolved images from real images. With this approach our generator can learn to create solutions that are highly similar to real images and thus difficult to classify by D. This encourages perceptually superior solutions residing in the subspace, the manifold, of natural images. This is in contrast to SR solutions obtained by minimizing pixel-wise error measurements, such as the MSE.

At the core of our very deep generator network G are B residual blocks with identical layout [4]. Specifically, we use two convolutional layers with small 3×3 kernels and 64 feature maps followed by batch-normalization layers [5] and Parametric ReLU [6] as the activation function. This constitutes of a single block that is replicated five times in the generator network. We increase the resolution of the input image with two trained sub-pixel convolution layers as proposed by Shi et al. [7].

To discriminate real HR images from generated SR samples we train a discriminator network. The architecture is shown in Figure 1. We follow the architectural guidelines summarized by Radford et al. [8] and use LeakyReLU activation ($\alpha = 0.2$) and avoid max-pooling throughout the network. It contains eight convolutional layers with an increasing number of 3×3 filter kernels, increasing by a factor of 2 from 64 to 512 kernels as in the VGG network. Strided convolutions are used to reduce the image resolution each time the number of features is doubled. The resulting 512 feature maps are followed by two dense layers and a final sigmoid activation function to obtain a probability for sample classification.

WRITE ABOUT LOSS

5. Experiment

5.1. Dataset

KAIST dataset consisting pixel-by-pixel RGB and thermal LR image pair are given as input to the modal to produce the HR image. Dataset was developed using imaging hardware consisting of a color camera, a thermal camera and a beam splitter to capture the aligned multispectral (RGB color + Thermal) images.

The KAIST Multispectral Pedestrian Dataset consists of 95k color-thermal pairs (640x480, 20Hz) taken from a vehicle. Dataset consists of 70,679 images taken from videos. To get visually different images for better training, one out of every 40 images were picked and used for training and testing.

5.2. Results

6. Future Scope

The idea of adding the fourth channel containing the thermal information can be modified by using several techniques described below:

(a) Using separate generators for RGB and Thermal images and fused the two network's output features to generate the super-resolved image.

(b) Using Cross Modal Transfer Learning based on Hybrid Transfer Networks.

(c) Using Neural Image Priors with randomly initialized ConvNets with Image structure as the Priors without any learning.

7. Conclusion

We have described a modified deep residual network SRResNet that makes use of a thermal image features to further increase the quality of the super-resolved image. We have highlighted some limitations of this PSNR-focused image super-resolution and introduced SRGAN, which augments the content loss function with an adversarial loss by training a GAN. We have confirmed that SRGAN reconstructions for large upscaling factors (4×) are more photo-realistic than reconstructions obtained with the existing methods.

8. References

[1] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint.

[2] C.-Y. Yang, C. Ma, and M.-H. Yang. Single-image super-resolution: A benchmark. In European Conference on Computer Vision (ECCV), pages 372–386. Springer, 2014. 1, 2

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems (NIPS), pages 2672–2680, 2014.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).Mousavi, H. S., Guo, T., & Monga, V. (2018). Deep Image Super Resolution via Natural Image Priors. arXiv preprint arXiv:1802.02721.

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of The 32nd International Conference on Machine Learning (ICML), pages 448–456, 2015.Sapiro, G., "Inpainting The Colors", IMA Preprint Series 1979, May 2004

[6] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In IEEE International Conference on Computer Vision (ICCV), pages 1026–1034, 2015.

[7] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1874–1883, 2016.

[8] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In International Conference on Learning Representations (ICLR), 2016

[9] Sharmila, T., & Leo, L. M. (2016, April). Image upscaling based convolutional neural network for better reconstruction quality. In *Communication and Signal Processing (ICCSP), 2016 International Conference on* (pp. 0710-0714). IEEE.

[10]