Better Image Super-Resolution Using Pixel-Wise Semantic Segmentations
CS 381V Visual Recognition Final Project

Keivaun Waugh
University of Texas at Austin
keivaunwaugh@gmail.com

Paul Choi
University of Texas at Austin
choipaul96@gmail.com

Abstract

In this paper, we address the problem of image super-resolution (SR) using a ResNet based convolutional neural network (CNN) as well as a generative adversarial network (GAN). Our approach differs from prior work with the addition of explicit semantic segmentation data. We experiment with both human annotated pixel-wise segmentations as well as machine generated pixel-wise segmentations. Our results show that adding the segmentation data to the network increases both quantitative and qualitative performance of the SR network.

1. Introduction

Image super-resolution is a compelling area of computer vision. Being able to convert low resolution photos and video to higher resolution counterparts is an important problem as hardware continues to improve. With higher resolution displays and cameras more available today than ever before, it would be nice if we could convert old media to the same resolution standards as current photos and video. While image downsizing is an easy problem with a transformation that can be easily specified, SR is an inherently underspecified problem. The goal is to create an inverse transformation that "hallucinates" information from the low resolution (LR) image to fill in the gaps for the high resolution (HR) image. Naive methods of upsampling using nearest neighbor and bicubic interpolation introduce many artifacts and produce visually displeasing images. Figure 1 displays an example of results obtained with a naive method as well as perfect results to which we aim to come closer.

Deep networks have greatly increased in popularity for instance and category recognition, sparked by work from Krizhevsky et al. [11]. SR methods based on deep networks have also achieved considerable success in SR tasks. We use the success of these deep networks as a starting point for our research. In this paper, we combine semantic segmentation features with state-of-the-art SR techniques in order to achieve sharper SR solutions. We explain the intuition behind adding this pixel-wise segmentation data later in this paper.

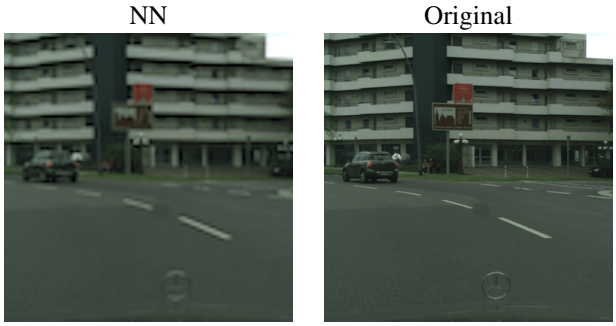

Figure 1: Left: image upscaling using nearest neighbor. Right: The original image before downscaling then upscaling.

1.1. Related Work

Before the advent of deep neural networks, basic approaches to SR were available. The most naive is nearest neighbor which increases the resolution with a simple pixel copy of the nearest pixel in between the LR and HR images. A more common approach that typically gets better results and is the standard in many image editors is bicubic interpolation. However, neither of these two methods attempt to use any kind of local or global structure in the image to perform a better reconstruction.

Other, more advanced non-deep approaches exist for SR. Chung et al. [4] investigate the use of fractal patterns for the task of super resolution. They were able to achieve better results than the naive methods. Approaches for SR that involve video like those mentioned in Borman and Stevenson's review [2]. These algorithms aim to use context of neighboring frames to pick values for the upsampled pixels. However, we choose to stick to the problem of single image SR.

Dong et al. [6] were the first to propose a CNN for the purpose of SR. They used a basic deep architecture

and trained their model end-to-end using a per-pixel loss between the output and ground truth images. Johnson et al. [9] proposed the use of a perceptual loss function for SR to make the images more appealing to eye. They get worse quantitative results such as peak signal-to-noise-ratio (PSNR) and structural similarity (SSIM), but better qualitative results. This helped to motivate future work to use perceptual metrics such as mean opinion score (MOS). Other similar work involving CNNs for SR include [17] and [10].

Another more recent approach is to use a generative adversarial network (GAN) [7]. As mentioned previously, the CNN approaches typically attempt to minimize the per-pixel loss between the upsampled output image and the original unmodified image. The error metric frequently used is peak signal-to-noise ratio (PSNR). However, when this is applied directly on the pixel space, this often encourages the network to make soft changes in the upsampled image rather than generate the high-frequency changes often found in real images. GAN based approaches like the work of Ledig et al. [12] attempt to overcome this by using an adversarial network. Though they get lower PSNR values, the images often look more realistic, which suggests that some other metric should be optimized to get better results.

In [3], Chen and Koltun experiment with using semantic segmentation from the Cityscapes dataset [5] to perform photorealistic image synthesis. The authors are able to get quote compelling results using only the semantic segmentations as input to their CNN. However, to our knowledge, there have been no papers published that attempt to merge together this explicit semantic data with a SR technique for better results.

1.2. Contribution

Semantic segmentation is a core focus of much computer vision work. In this paper, we are the first to explore the how the explicit addition of pixel-wise semantic segmentation masks affect an SR network's ability to produce quality results. We experiment with a high upscaling factor ($4\times$). Our hypothesis was that the segmentations will clean up ambiguous edges, especially at high upscalings when the borders are especially fuzzy. We also evaluate the effect of these segmentations on both more traditional ConvNet architectures as well as GANs that have become more recently popular for the image generation domain.

We describe our network architecture and segmentation pipeline in section 2. A quantitative evaluation of results is included in section 3. Section 10 contains concluding remarks and a direction for future work based on our segmentation approach.

2. Method

Our solution to this task involves using a standard convolutional neural network based on the ResNet architecture [8] as well as a generative adversarial network [7]. Both of these use modified architectures that were designed by the authors of [12]. We modified the input layers of the generator network to take in the extra information supplied by the segmentation pipeline. The two variants of this pipeline will be discussed in future sections. When running the network in the standard ResNet mode, only the generator is trained. When running in GAN mode, the generator is first pretrained for n epochs, followed by m epochs of joint training between the generator and the discriminator. Typically $n < m$, but this is not a hard requirement.

2.1. Architecture

Figure 2 illustrates the architectures for both the generator and the discriminator.

2.2. Human Segmentations

All the datasets that we evaluated our system on contain human annotated segmentations. These datasets are [5] and [13]. Due to the large performance gap between current segmentation networks and human annotated results, we wanted to evaluate our method in the best case possible as a sanity test. If we were unable to get better results with the human segemntations, then it would be unlikely that we were get better performance with the machine segmentations at the present time, or even as they get better. Not until they became significantly better than humans would there be a chance of performance gains.

2.3. Machine Segmentations

Talk about reasons for using FCN [14] over some other solution that gets better performance. Discuss how it wasn't actually all that crucial to get great performance out of the machine segmentations after we showed that human segmentations get gains over no segmentations.

3. Evaluation

While discussing this problem with our peers, some have argued that if having semantic segmentations added to the network would increase performance, the vanilla GAN in [12] should learn how to segment items in the scene "under the hood". We argued this was not true and pointed to [8] as an example. The authors of this paper showed that it was difficult for deep networks to learn identity mappings as the depth of networks increased. We believed that a similar phenominon would occur in this domain and that adding the segmentation information explicity, we could overcome this learning difficulty. Our evaluations show that this is the case.

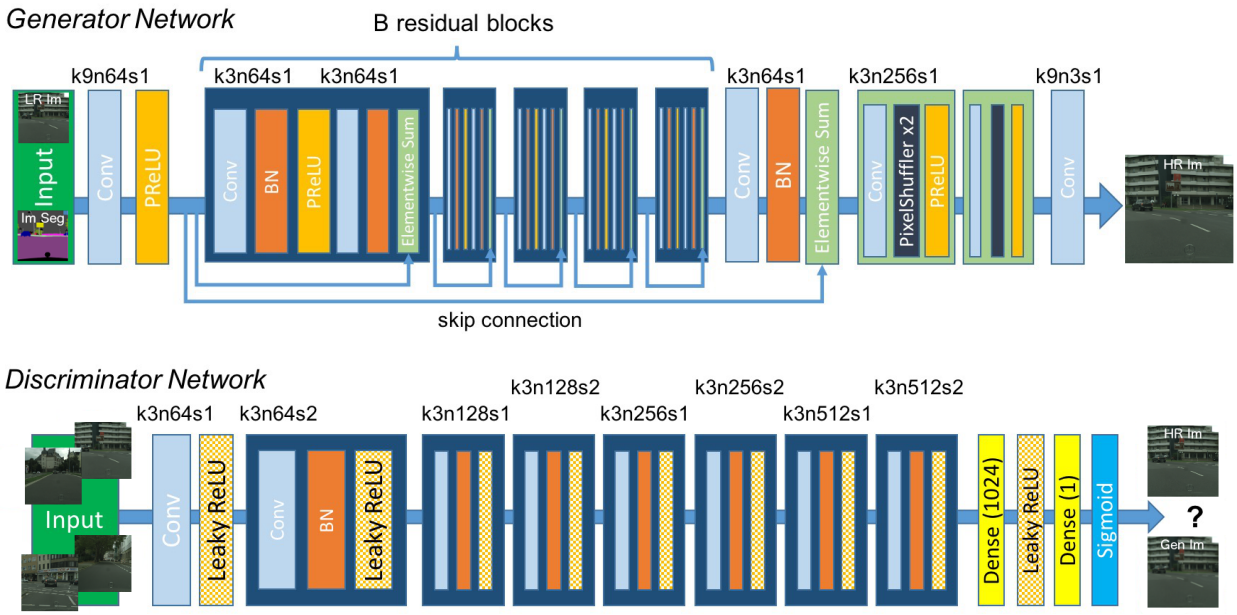

Figure 2: Architecture of Generator and Discriminator Network with kernel size, feature map count, and stride at each layer in the network. Diagrams courtesy of [12] as this was the basis of our architecture.

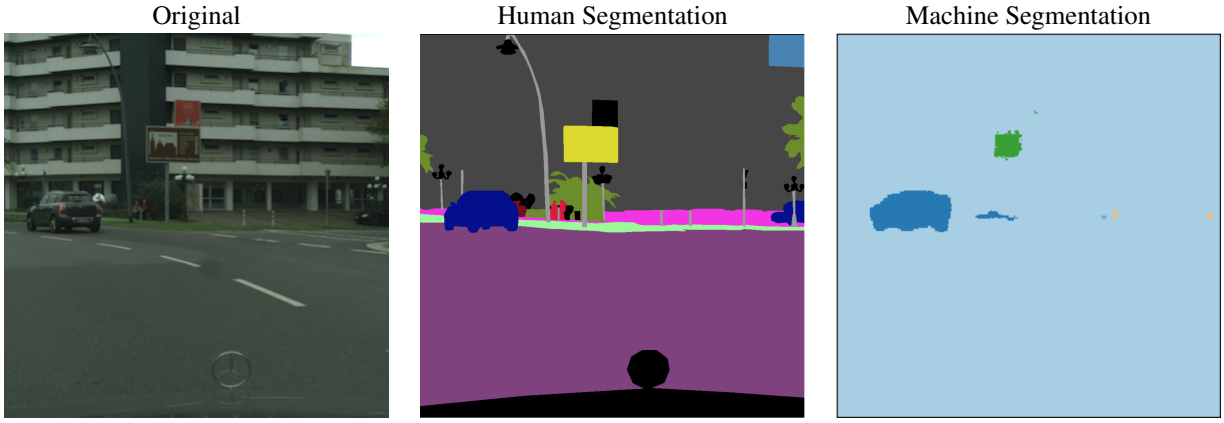

Figure 3: Comparison of segmentations between human generated and auto generated. Left: the original image. Middle: A human annotated pixel-wise semantic segmentation. Right: A machine generated segmentation.

As is standard in other papers in this domain such as [6], [12], [9], [10], we use the metrics of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) as quantitative measurements. However, as shown in [12], quantitative measurements do not always translate to increases in perceptual qualtiy. As a result, we have decided to adopt the use of a mean opinion score (MOS) and have employed the use of our peers to help us gather data for this metric.

Filler

4. Technical Plan

The authors of [12] have their code for their SRGAN network available online. We will use this code as a starting point and modify it to include our semantic segmentation data from the datasets or out-of-the-box segmentation networks. One method we will use for integration this additional data will be to concatenate the segmentation data onto the feature maps of the early layers in the network. This will be evaluated on both the traditional ResNet variant that the

authors introduce as well as the GAN version.

5. Experimental Plan

We will build upon the open-source implementation of Ledig et al. [12] which uses TensorFlow 1.2. Our first task will be to apply the unmodified network to a dataset of our choosing which includes semantic segmentation labels. These results will constitute our baseline. We will use the same measures to evaluate our SR method: PSNR, structural similarity (SSIM), and mean opinion score (MOS). For MOS, we will conduct a survey as in [12] and ask raters to rate images produced by different methods on a scale of 1 to 5 in terms of quality. While the original paper uses 26 raters, we will likely use fewer raters to ease the logistical burden. As an additional baseline, we will also measure the performances of the nearest neighbor and bicubic interpolation methods.

Our second and main task will be to modify the SRResNet and SRGAN networks to utilize semantic segmentation features, as discussed in the technical plan. If we have enough time, we will also experiment with adding these additional features to the discriminator in the SRGAN network to see if it aids training. The measurements from these modified approaches will indicate the effectiveness of our method.

In addition to using a dataset with semantic segmentation labels, we will also use datasets without such labels included and generate the labels manually using an out-of-the-box solution. This will allow us to compare how human-annotated segmentation features compare to machine-generated segmentation features.

We will experiment with adding the segmentation data at multiple places in the network. The most basic solution would be to add it onto the feature maps at the earliest layers in the network. However, we plan to evaluate the network with the segmentations included at different locations and multiple locations.

6. Sources of Data

Image SR techniques are self-supervised in that explicit annotation of data is unneeded. Given an image, a training pair can easily be generated by using the downsampled image as the input to the SR network and the original image as the ground truth. However, in order to integrate semantic segmentation knowledge, we need to obtain the semantic layout for each image. There are two ways that we plan on going about this. The first is to use datasets that already include the segmentation such as the Cityscapes [5] dataset. The other approach is to use existing segmentation networks as in [14] to generate the segmentation before the downsized image is fed into our SR network. We plan to compare the results of the two to see if the more accurate human-annotated segmentations are required to get good results from our technique, or if a fully-automated SR approach can be achieved.

To compare our results against those obtained from other papers, we plan on using standard evaluation datasets that are used in [12]. These include Set5 [1], Set14 [18], and BSD100 which is a subset of BSD300 [15]

7. Partner Plan

For the programming portion of the project, we plan to pair-program and divide equally the implementation work. For experiments, we will individually drive each one while working together as needed, with each person driving an equal number of experiments. The report will be written collaboratively. The person with more expertise on a given section will lead writing for that section.

8. Speculation of Results

Current state-of-the-art semantic segmentation methods already involve CNNs. One may argue that a CNN based approach for SR would automatically learn information regarding segmentation if it is useful for providing better results. However, this argument does not take into account the difficulty of learning such information. A parallel argument was made for residual networks in [8] with regards to learning identity mappings. However, the authors found that it was often difficult for networks to learn such mappings. We predict that adding the semantic segmentation data explicitly to the network will allow it to make more intelligent decisions along object boundaries. We expect this will help provide crisper edges. Because of the success with using semantic masks at various resolutions in [3], we expect that our results will be better at high upsampling factors.

9. Experiments

Filler

10. Conclusion and Future Work

10.1. Omissions from Project Proposal

I think in this section we should include a discussion of stuff that we said we would do in the proposal that we didn't get around to doing in the final project. We should justify why these were left out (whether it was because we ran out of time, didn't think it would be that useful/sightful, etc.)

Filler

References

[1] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Alberi-Morel. Low-complexity single-image super-resolution based

on nonnegative neighbor embedding. In *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, pages 1–10, 2012.

[2] S. Borman and R. Stevenson. Spatial resolution enhancement of low-resolution image sequences - a comprehensive review with directions for future research. Technical report, 1998.

[3] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. *CoRR*, abs/1707.09405, 2017.

[4] K.-H. Chung, Y.-H. Fung, and Y.-H. Chan. Image enlargement using fractal. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP 03).*, Jun 2003.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016.

[6] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):295–307, Feb. 2016.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[9] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016.

[10] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. *CoRR*, abs/1511.04491, 2015.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[12] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.

[13] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[14] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.

[15] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *in Proc. 8th Intl Conf. Computer Vision*, pages 416–423, 2001.

[16] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1605.06211, 2016.

[17] W. Shi, J. Caballero, F. Huszár, 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. *CoRR*, abs/1609.05158, 2016.

[18] R. Zeyde, M. Elad, and M. Protter. *On Single Image Scale-Up Using Sparse-Representations*, pages 711–730. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.