# CS 381V Visual Recognition Final Project Proposal

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

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

## Abstract

*In this project, we plan to explore the benefits of providing semantic segmentation data to an image super-resolution (SR) system.*

## 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. Naive methods of upsampling using bicubic interpolation introduce many artifacts and produce visually displeasing images. Deep networks have greatly increased in popularity for instance and category recognition, sparked by work from Krizhevsky et al. [8]. SR methods based on deep networks have also achieved considerable success in SR tasks. We plan to combine semantic segmentation features with state-of-the-art SR techniques in order to achieve sharper SR solutions.

## 2. Related Work

Prior work involving image synthesis with deep networks involves one of two techniques: using a more conventional CNN for upsampling [6] [12] [7] or using a generative adversarial network (GAN) [4]. Typically, the CNN approaches 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. [9] 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 [2], Chen and Koltun experiment with using semantic segmentation from the Cityscapes dataset [3] to perform photorealistic image synthesis. 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. We believe that this will provide for better results than what is currently achievable with the published methods.

## 3. Technical Plan

The authors of [9] 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.

## 4. Experimental Plan

We will build upon the open-source implementation of Ledig et al. [9] 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 [9] 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.

## 5. 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 [3] dataset. The other approach is to use existing segmentation networks as in [10] 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 [9]. These include Set5 [1], Set14 [13], and BSD100 which is a subset of BSD300 [11]

## 6. 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.

## 7. 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 [5] 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 [2], we expect that our results will be better at high upsampling factors.

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

[3] 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.

[4] 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.

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

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

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

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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