21 Nov 2016

Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi

1. Overview

The highly challenging task of estimating a high-resolution (HR) image from its low-resolution (LR) counterpart is referred to as super-resolution (SR).

Wide range of applications in areas such as:

- HDTV
- Medical imaging
- Satellite imaging
- Face recognition
- Surveillance

One central problem remains largely unsolved...

There have been several breakthroughs in accuracy and speed of single image super-resolution using faster and deeper convolutional neural networks, **but how do we recover the finer texture details when we super-resolve at large upscaling factors?**

Previous work has largely focused on minimizing the mean squared reconstruction error (the resulting estimates having high peak signal-to-noise ratios), but the estimates often lack high-frequency details and are perceptually unsatisfying (the images look blurry, i.e. images have overly smooth textures).

2. Proposed Solution

SRGAN

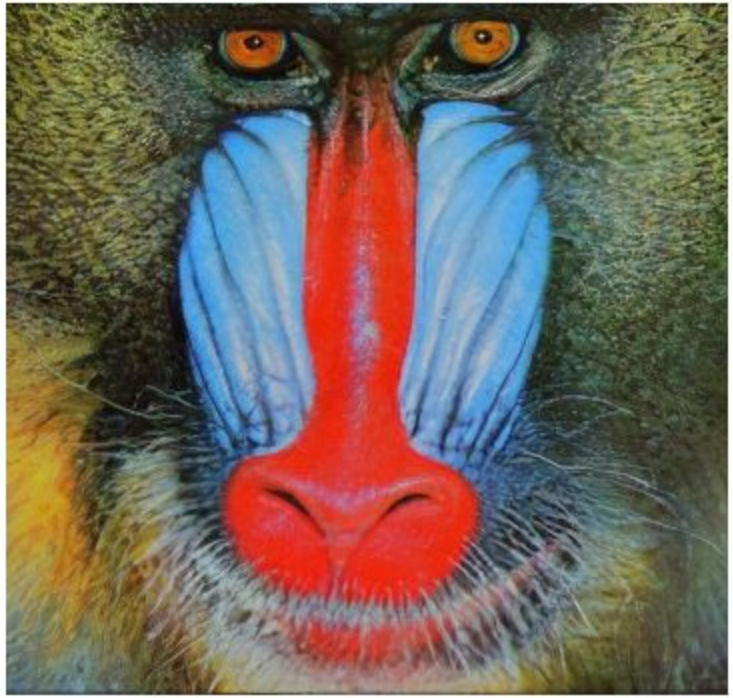
- The first framework capable of inferring photo-realistic natural images for 4x upscaling factors, utilizing a generative adversarial network.

To achieve this:

Perceptual loss function

- Adversarial loss
- Content loss

4× SRGAN (proposed)



original

3. Related Work

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. Minimizing MSE also maximizes the peak signal-to-noise ratio (PSNR), which is a common measure used to evaluate and compare SR algorithms. 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.

This paper describes the first very deep ResNet architecture using the concept of GANs to form a perceptual loss function for photo-realistic SISR.

bicubic
(21.59dB/0.6423)

SRResNet
(23.53dB/0.7832)

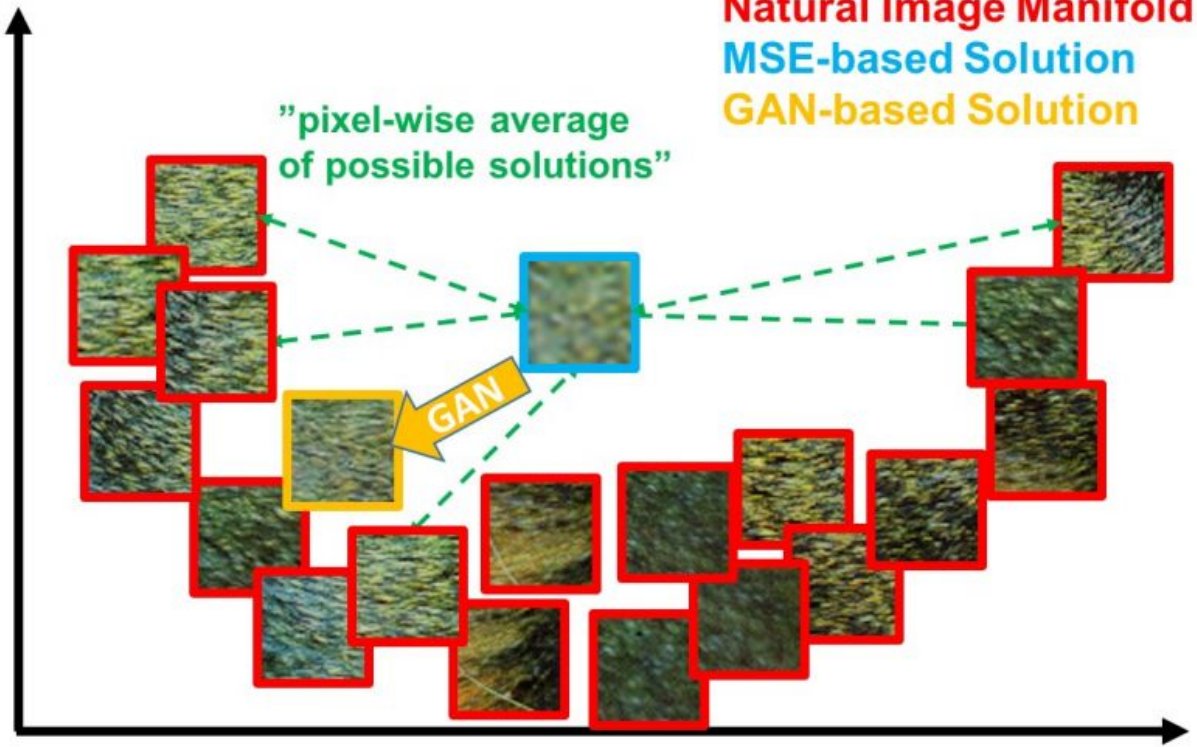
SRGAN
(21.15dB/0.6868)

original

Highest PSNR does not necessarily reflect the perceptually better SR result.

(Corresponding PSNR and SSIM are shown in brackets)

The MSE-based solution appears overly smooth due to the pixel-wise average of possible solutions in the pixel space, while GAN drives the reconstruction towards the natural image manifold producing perceptually more convincing solutions.



4. Contributions

- The authors set a new state-of-the-art for image SR with high upscaling factors as measured by PSNR and structural similarity (SSIM) with their 16 blocks deep ResNet (SRResNet) optimized for MSE.
- SRGAN, which is a GAN-based network optimized for a new perceptual loss. Here the authors replace the MSE-based content loss with a loss calculated on feature maps of the VGG network, which are more invariant to changes in pixel space.
- The authors confirm with an extensive mean opinion score (MOS) test on images from three public benchmark datasets that SRGAN is the new state of the art, by a large margin, for the estimation of photo-realistic SR images with high scale factors (4x).

5. Method

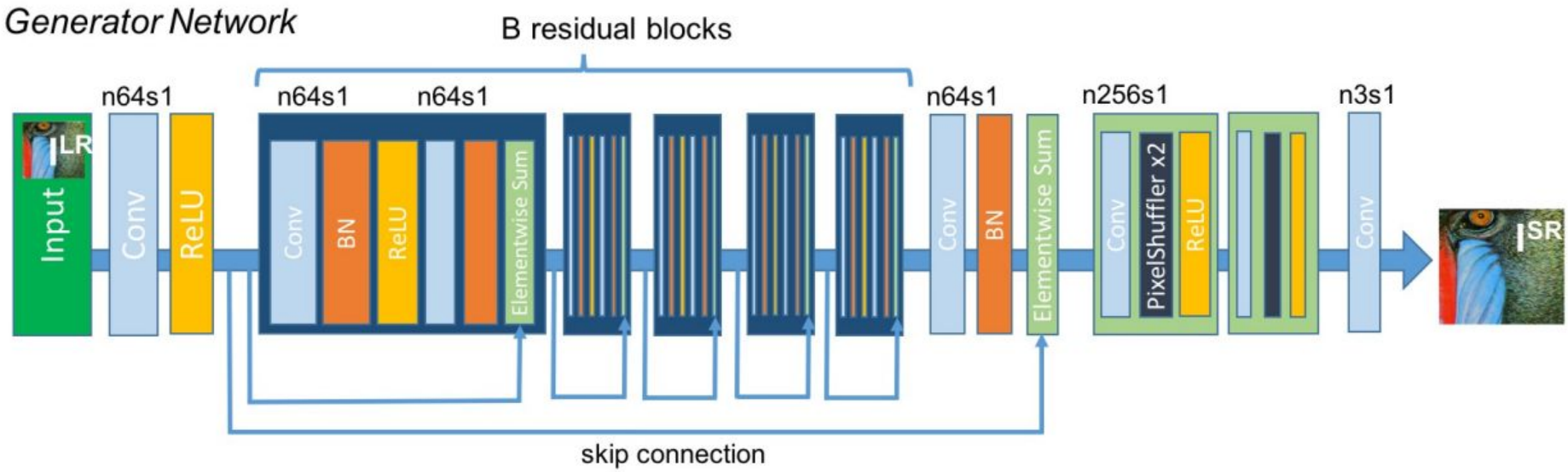
In SISR the aim is to estimate a HR, super-resolved image from a LR input image. The LR image is taken from its HR counterpart, and the HR images are only available during training. In training, LR images are obtained by applying a Gaussian filter to their HR images followed by a downsampling operation with downsampling factor r.

The ultimate goal is to train a generating function G that estimates for a given LR input image its corresponding HR counterpart. To achieve this, the authors train a generator network as a feed-forward CNN G_θG.

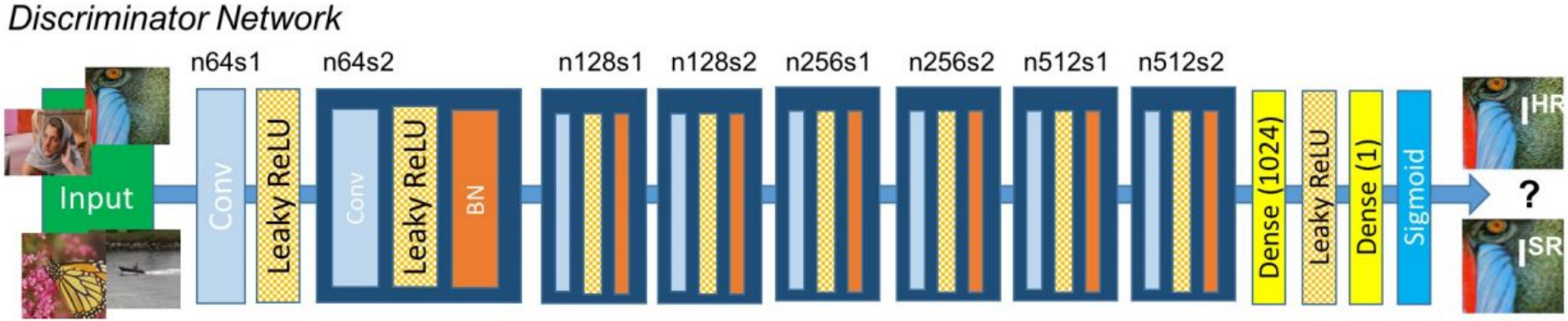
Adversarial Network Architecture

A discriminator network $D_{\theta D}$ is further defined.

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 the generator can learn to create solutions that are highly similar to real images and thus difficult to classify by D.



Two convolutional layers are used, with small 3 x 3 kernels and 64 feature maps followed by batch-normalization layers and ReLU as the activation function. Resolution of the input image is increased with two trained sub-pixel convolution layers as proposed by Shi et al.



To discriminate real HR images from generated SR samples, the authors trained a discriminator network. It contains eight convolutional layers with an increasing number of 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.

Perceptual Loss Function

Content Loss

- Motivated by perceptual similarity instead of similarity in pixel space such as the pixel-wise MSE loss (the most widely used optimization target).

Adversarial Loss

- The adversarial loss pushes solutions to the natural image manifold using a discriminator network that is trained to differentiate between the super-resolved images and original photo-realistic images.

6. Experiments

Experiments were performed on three widely used benchmark datasets Set5, Set14, and BSD100. All experiments are performed with a scale factor of 4x between low- and high-resolution images. This corresponds to a 16x reduction in image pixels.

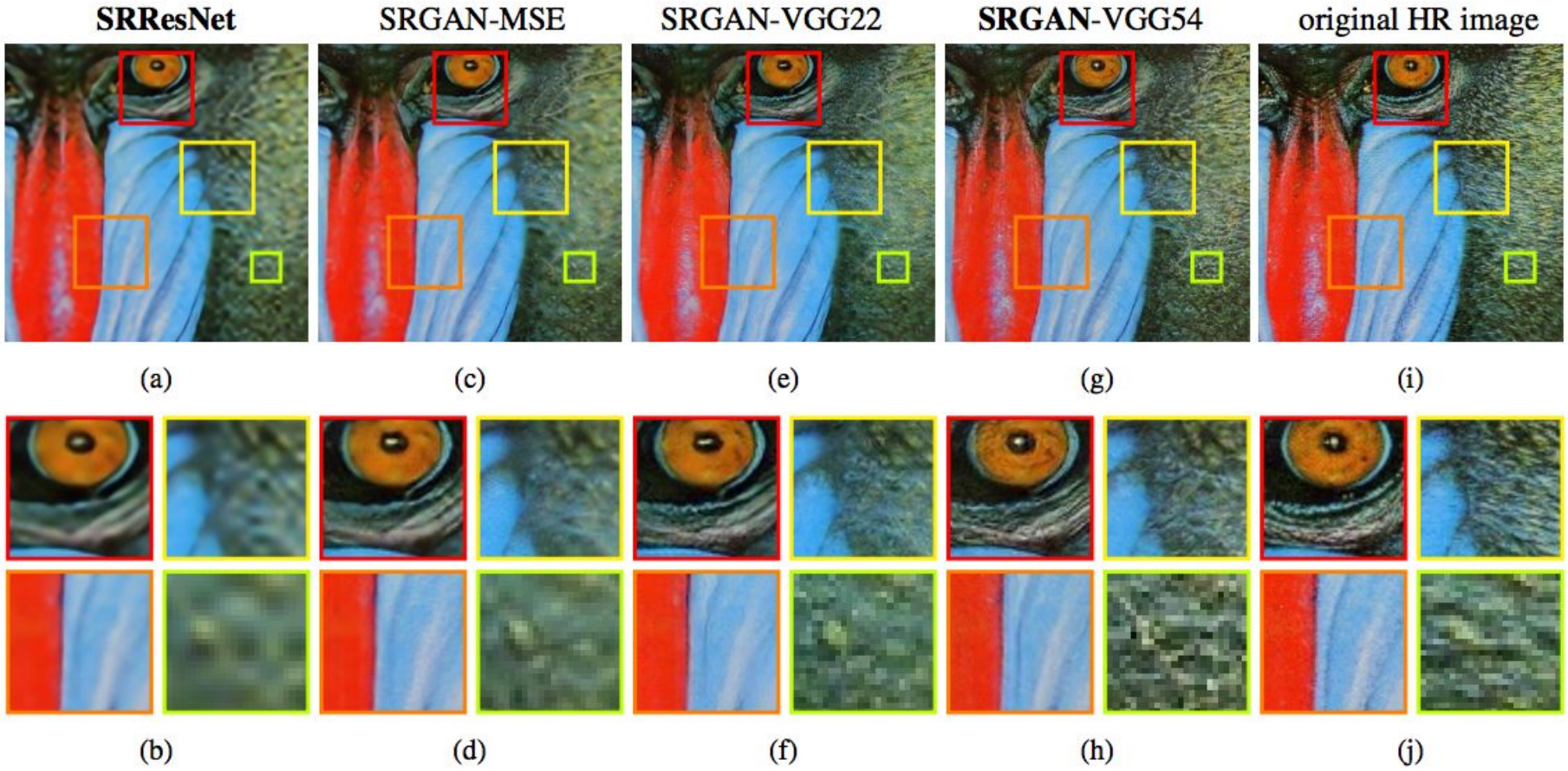
All networks were trained on a NVIDIA Tesla M40 GPU using a random sample of 350 thousand images from the ImageNet database. These images are distinct from the testing images.

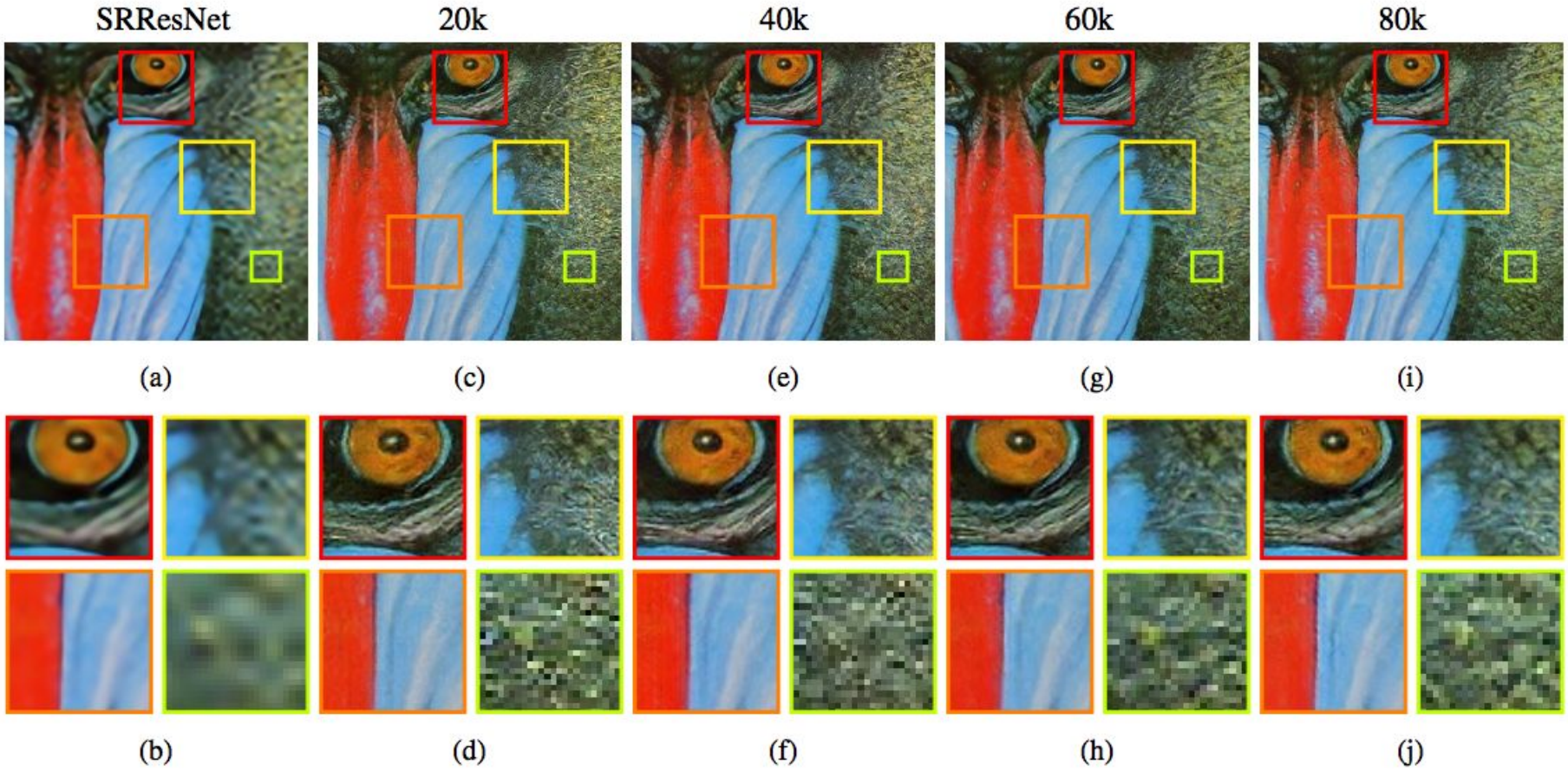
Note: the generator model can be applied to images of arbitrary size as it is fully convolutional.

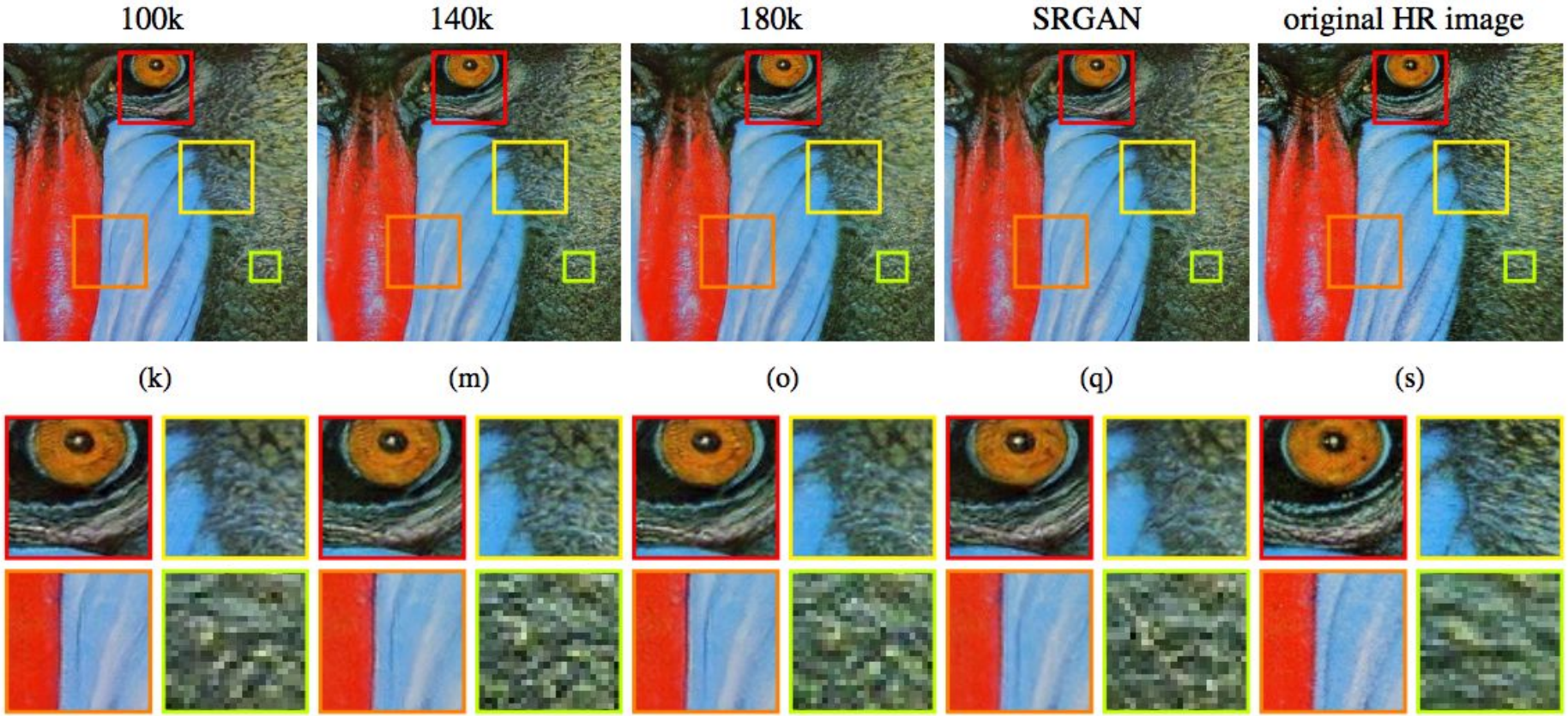
All SRGAN variants were trained with 10^5 update iterations at a learning rate of 10^−4 and another 10^5 iterations at a lower rate of 10^−5. Updates to the generator and discriminator network alternate. The generator network has 16 identical (B = 16) residual blocks. During test time, batch-normalization update is turned off to obtain an output that deterministically depends only on the input. Final implementation is based on Theano and Lasagne.

Conclusion

- Standard quantitative measures such as PSNR and SSIM fail to capture and accurately assess image quality with respect to the human visual system.
- The focus of this work was the perceptual quality of super-resolved images rather than computational efficiency.
- The presented model is not optimized for video SR in real-time.
- Of particular importance when aiming for photo-realistic solutions to the SR problem is the choice of the content loss.
- The ideal loss function depends on the application. For example, approaches that hallucinate finer detail might be perceptually convincing but less suited for medical imaging or surveillance.


SRResNet
SRGAN-MSE
SRGAN-VGG22
SRGAN-VGG54
original HR image
(a)
(c)
(e)
(g)
(i)
(b)
(d)
(f)
(h)
(j)


SRResNet
20k
40k
60k
80k
(a)
(c)
(e)
(g)
(i)
(b)
(d)
(f)
(h)
(j)


100k
140k
180k
SRGAN
original HR image
(k)
(m)
(o)
(q)
(s)
(l)
(n)
(p)
(r)
(t)


bicubic
SRResNet
SRGAN
original

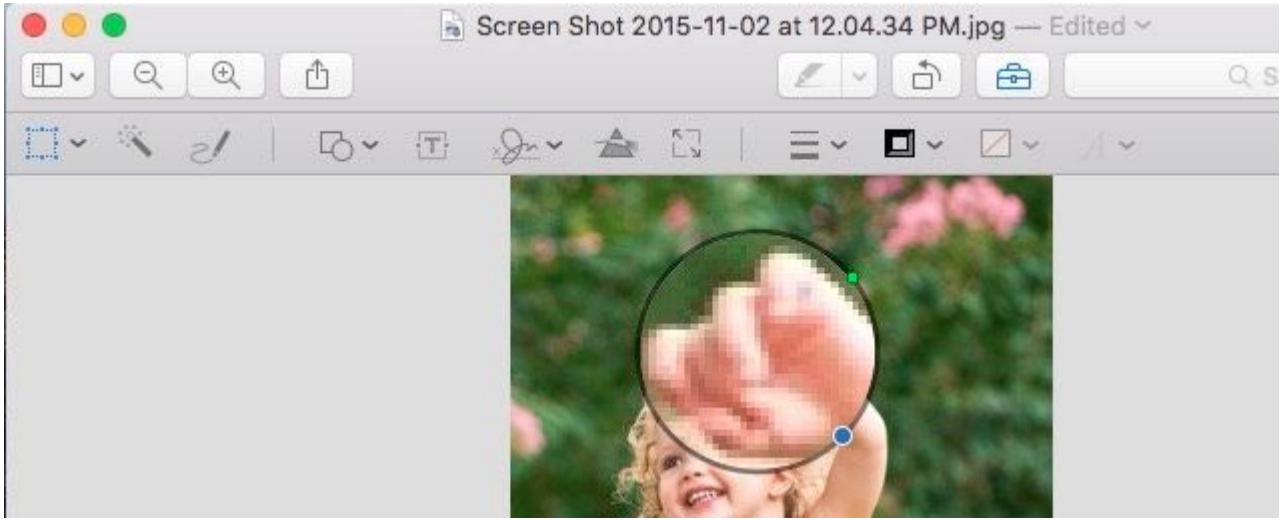







Final Project Proposal

SRGAN-Magnifier



NOTE: Provided as reference for general idea of final project proposal. Instead of magnifying the original LR photo, which gets pixelated, our application will use the image generated by the SRGAN in the magnifier view.