DTU Compute
Department of Applied Mathematics and Computer Science

Super-Resolution using a SRGAN for Retinal Ophthalmoscopy

Matthias Franz-Josef Rötzer

1 Introduction

The challenge of upsampling of images has been an important topic in digital image processing since its very beginning. Especially with the continuous development of deep learning frameworks, Single-Image Super-Resolution (SI-SR) gained more popularity and has a mainly been applied in domains where undersampled or compressed data is prevalent. In this poster, we explore the capabilities of those methodologies on a special set of three channel images.

2 Dataset

As our training data, we chose a set of high-resolution 'Fundus' Images, originating from the FAU in Germany [1]. These images provide a detailed view of the retina, which is located at the back of the eye and plays a crucial role in vision. To allow for more memory-efficient training, the images have been sliced into 512 by 512 pixel wide sub-images (patches). We are producing linescan-like input training images by sampling every 4th pixel in the x- and y-direction (see Fig. 2, left). Because of the cropping that will later be done in the contracting convolutional layers, and we also want to predict the pixels in the border region, we are increasing the image size by mirroring [2].

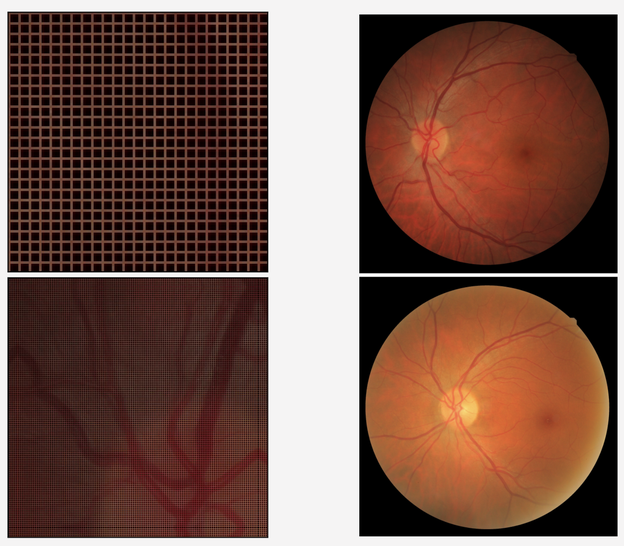

Figure 1: 'Linescan' input training data (left), Example of the ground truth input (right)

3 UNet

U-Net is a CNN architecture which is composed of symmetric contracting (encoder) and expanding (decoder) paths, with layer-wise copy connections between them. U-Nets are commonly used for image segmentation tasks, but by doing some adjustments to the output layer and channels, it is also possible to apply it for single-image super-resolution tasks. We will include it into our comparison to the performance of the Generative Adversarial Network (GAN) (see Section 6 and Table 1).

4 SRGAN

Convolutional neural networks have been used to generate high-resolution images that train faster and achieve high accuracy. However, in our case, and so in some other cases, they are unable to recover high-frequency (finer) details and often produce blurry images. The now discussed Super-Resolution GAN (SRGAN) architecture combats most of these problems to produce realistic looking images with high perceptual quality. As proposed by Goodfellow et al. in 2014 [3], the GAN architecture can be broken down into two adversarial, in other words 'competing' models: the generator and the discriminator. We have already discussed a UNet image generation model, which will now present itself as the generator model. By adding the discriminator model, which acts as an image classifier, we want to ensure that the overall architecture adapts appropriately to the quality of the images and the resulting images are much more optimal than a single CNN upsampled reconstruction.
For the discriminator, we make use of the VGG-16 model as suggested by [4] - even though they suggest using it as a so-called perceptual loss. We're using the *torchvision* hosted pretrained weights and change the classifier, so that it consists of three layers of linear transformation, ReLU activation, and dropout operations, followed by one final linear layer with a sigmoid activation function. Generally in SRGANs we distinguish between two loss operations; A content-loss which is to measure how similar to the ground truth (real image) an image is and the adversarial-loss, which ensures that the generated image is realistic i.e. not easily distinguishable from the real images. The latter will prove to be challenging to balance.

5 Evaluation

To evaluate the objective image quality, assessment metrics such as the peak signal-to-noise ratio (PSNR) in connection with the mean-squared-error (MSE) is commonly used. The MSE represents the cumulative squared error between the generated I^{SR} and the original image I^{HR}, whereas $PSNR$ represents a measure of the peak error. The lower the value of MSE, the lower the error.

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I^{SR}(i,j) - I^{HR}(i,j)]^2$$

$$PSNR = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_{I^{HR}}^2}{\mathrm{MSE}}\right)$$
$$= 20 \cdot \log_{10}(\mathrm{MAX}_{I^{HR}}) - 10 \cdot \log_{10}(\mathrm{MSE})$$

$SSIM$ is another method of predicting perceived quality and will be used to measure the similarity between two images. If this value is close to 1, two images are identical. Otherwise, two images will be completely different.

	Lin. Baseline	U-Net	SRGAN
SSIM	0.89	0.84	0.86
MSE	7.39	97.35	92.36
PSNR	370.88	362.81	363.26

Table 1: Pixel-wise similarity measures

6 Results

Table 1 shows the mean SSIM, MSE and PSNR over the entire dataset (1920 sub-images). Unfortunately, our current SRGANs setup doesn't seem to show significant differences to a U-Net in terms of objective similarity to the ground truth. Speaking in terms of visual subjective inspection, there are slight improvements, but still not as good as our linear baseline (see Fig. 6). This could be mainly due to the instability of the adversarial loss during training and the lack of more complex losses [4], [5]. Finer details, that are visible in the ground truth image, don't seem to emerge from our. An additional aspect to improve upon is the overall optimization of the runtime, both in terms of data load and duration, since we are essentially training two networks and calculating multiple losses. An optimization would allow us for more extensive hyperparameter searches and cross-validation.

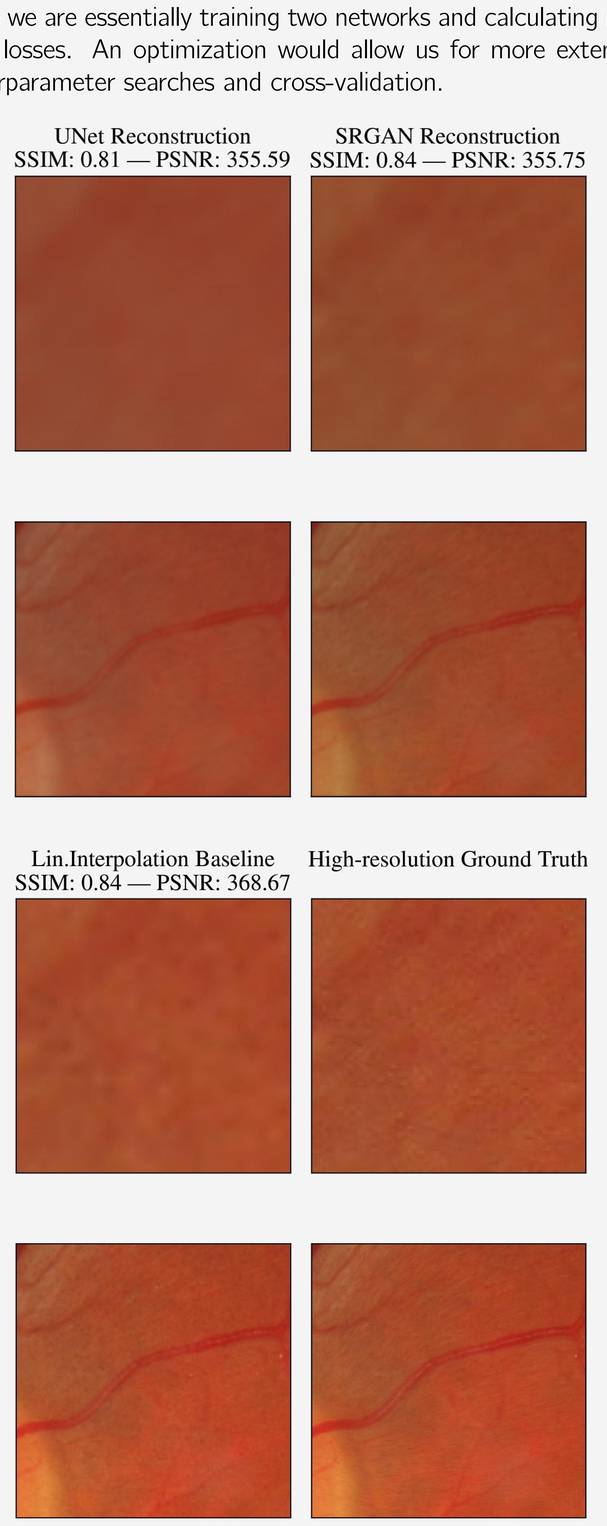

Figure 2: Comparison example of the generated out

7 Outlook

There's of course still room for improvement. One possibility is adding further measures of loss like perceptual losses, which measure the difference between the high-resolution output of the generator and the ground truth high-resolution image in terms of perceptual features such as texture, color, and sharpness, rather than just pixel values. Perceptual losses can be computed using pre-trained neural networks that have been trained on large image datasets, such as VGG or ResNet. By comparing the feature maps of the generated and ground truth images at different layers of these networks, we can obtain a measure of perceptual similarity [4].

References

[1] "High-resolution fundung images, fau, germany." `https://www5.cs.fau.de/research/data/fundus-images/`. Accessed: 2023-04-30.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[4] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016.

[5] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "Esrgan: Enhanced super-resolution generative adversarial networks," 2018.

DTU Compute
Department of Applied Mathematics and Computer Science
DTU Compute
Department of Applied Mathematics and Computer Science
DTU Compute
Department of Applied Mathematics and Computer Science
DTU Compute
Department of Applied Mathematics and Computer Science