Final project report
SinGAN: Learning a generative model from a Single Natural Image

RAMI Hamza hamza.rami@student.ecp.fr

Moukadiri Amine mohammed-amine.moukadiri@student.ecp.fr

Abstract

This document is the final project report that aims to experiment with SinGAN and propose an extension in term of both model design and also the training. The approach consists on adding a additional contextual loss to SinGAN model and apply it for image inpainting, and specifically for subtitles removing. "User studies" confirm that the generated patches (previously the text masks in the images) are commonly confused to be real images.

1 Introduction

Image inpainting, usually refers to filling missing pixels of an image, is an important task in computer vision. It has many applications such the one we are interested in which is subtitles removing and inpainting the missing text part. With the rapid progress in deep convolutional neural networks (CNN) and generative adversarial networks (GAN), many recent works have shown great results. We will be using SinGAN model with a custom loss function for its ability to train in a single image and generate random images at different scales. We will present in the next sections the hole architecture of SinGANs and also we will propose and develop our approach to solve the subtitles removing using SinGANs with a modified loss function, and finally we will evaluate our results in both qualitative and quantitative terms and compare it with existing approach (DeepFill v2)

2 SinGAN

Unlike many other inpainting and GANs based models, SinGAN (Tamar Rott Shaham, 2019) is trained only on a single image. It is based on the assumption that the internal statistics of patches within a single image typically carry enough information for learning a powerful generative model. This explains the architecture of SinGAN, as you can see bellow:

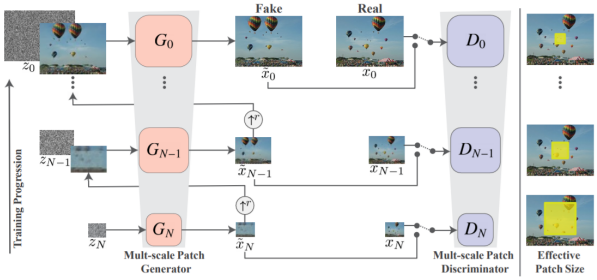

Figure 1: SinGANs multi-scale pipeline ,to see the official implementation please refer to https://github.com/tamarott/SinGAN

2.1 Contextual loss

In the original model we had a loss composed of Adversarial Loss and a reconstruction loss. The adversarial loss L_{adv} penalizes for the distance between the distribution of patches in x_n where x_n is a downsampled version of the input image by a factor r^n, for some r > 1 and the distribution of patches in generated samples $\tilde{x}_n$. The reconstruction loss L_{rec} insures the existence of a specific set of noise maps that can produce x_n, In the inpainting model we added a contextual loss to help the model to focus more in the image outside the area of the subtitles. To do so we added a contextual loss for each GAN, which is the mean of the difference between the downsampled original image multiplied by the mask and the generated image multiplied by the mask. So the formula of this loss is :

$$L_{contextual} = mean(x_n.mask - \tilde{x}_n.mask)$$

2.2 Inpainting extension

Inpainting is the name given to the technique of reconstructing deteriorated images or filling in missing parts of an image.
In our case we decided to apply inpainting for the removal of subtitles in an image, but the approach could be generalized to any inpainting task.

2.2.1 detecting text

Figure 2: Initial image with subtitles

To detect the rectangular box containing the text we have used a simple approach. We started by converting the image to gray scale and then we applied a thresholding (simple binary threshold, with a handpicked value of 150 as the threshold value). And then we applied a dilation to thicken lines in image, leading to more compact objects and less white space fragments. We finally identified contours of objects in resulted image using opencv findContours function and drew a bounding box (rectangle) circumscribing each contoured object - each of them frames a block of text.

Figure 3: Text box detected and filled by the mean of neighbor pixels

To detect only the text, We have tried a basic approach which is a thresholding to keep only the white color because the subtitles were using are always white. We then fill the detected area with a mean of neighbors of each pixel.

Figure 4: mask of the rectangular box

Figure 5: output text detection

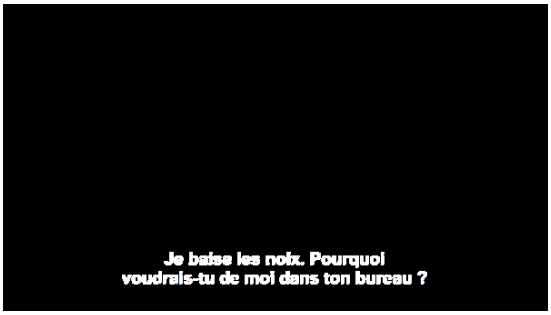

Figure 6: mask of the text

2.2.2 Training the SinGAN

We started by training the modified SinGAN on the original image containing the subtitles. and the mask that we generated in the previous step. We then applied **Harmonization**, it blend a pasted object with a background image. We train SinGAN on the background image, and inject a downsampled version of the naively pasted composite at test time. Preserving the objects structure and transferring the backgrounds texture.

2.3 Qualitative analysis

For the qualitative analysis we decided to send a questionnaire in the group of our

school we got 149 evaluation. The aim of this questionnaire was to give a score between 0 and 5 representative of how close the picture is to reality. We got the following histogram:

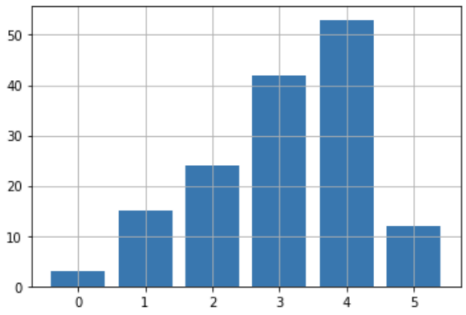

Figure 7: histogram notation

Figure 8: from left to right: a) original image, b) Inpainting output using the modified SinGan start scale = 3, c) start scale = 4

3 Comparison with DeepFill v2

3.1 DeepFill v2

DeepFill model (Jiahui Yu, 2018) is GAN based model composed by the coarse network which is trained with reconstruction loss explicitly, while the refinement network is trained with reconstruction loss, global and local WGAN-GP adversarial loss.

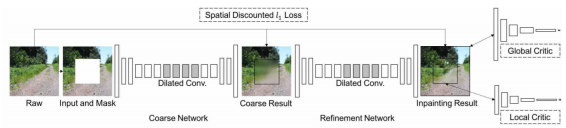

Figure 9: DeepFill architecture

For the v2 of the DeepFill model, based on coarse result from the first encoder-decoder network, a contextual attention layer is introduced in parallel with the dilated Conv and then merged to single decoder to get inpainting result.

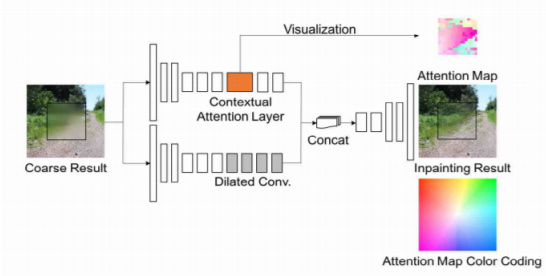

Figure 10: DeepFill v2 architecture

3.2 Comparison

Figure 11: From left to right: a) Initial image, b) Output with SinGAN, c) Output with DeepFill v2

The figure 11 shows that DeepFill shows really encouraging results regarding the inpainting with outputs that keep the same scale as the input image. On the other hand, SinGAN's approach generally outperforms DeepFill because it is a GAN based that model can learn from a single image, while with DeepFill we used the pretrained network to fill the text mask.

4 Conclusion

The proposed work that we have done, includes also another approach that shows promising results. The main idea is to fill the text mask with a pretrained DeepFill v2 and then do the Harmonisation of the generated box with the original image in order to make disappear the subtitles. For quantitative comparison, the Frechet Inception Distance (FID) (Martin Heusel, 2017) can be used to compare the real and fake image except that for our case, the original (real) image contains subti-

tles while the output (fake) image does not, which makes the comparison meaningless.

References

Jimei Yang Xiaohui Shen Xin Lu Thomas S. Huang Jiahui Yu, Zhe Lin. 2018. Generative image inpainting with contextual attention.

Thomas Unterthiner Bernhard Nessler Sepp Hochreiter Martin Heusel, Hubert Ramsauer. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium.

Tomer Michaeli Tamar Rott Shaham, Tali Dekel. 2019. Singan: Learning a generative model from a single natural image.