SinGAN, an inpainting extension

David Boudin
ENS Paris-Saclay
davboudin@wanadoo.fr

Pierre Ebert
ENS Paris-Saclay, University of Paris
p.ebert@hotmail.com

Abstract

Based on SinGAN[5], a single-image generative adversarial network, we implemented a model performing the task of inpainting. We applied our model to two sub-tasks of inpainting: recovering small defects and removing human silhouettes from images. Quantitatively comparing our inpainting model to a state-of-the-art inpainting tool [4] did not yield statistically significant results, due to small sample sizes and unreliable metrics. We published a qualitative questionnaire to measure how well our model could deceive the human eye. This survey revealed that our model performed almost as well as Nvidia's and can be very realistic, depending on the type of inpainting task faced.

1. Introduction

1.1. The inpainting problem

Within the realm of image manipulation techniques, image inpainting and image hallucination are important tasks. Image inpainting can be defined as the filling of small missing areas in an image by synthesizing new content, generally exploiting neighbouring information. Image hallucination is the filling of larger missing areas with synthetic content which requires the invention of plausible semantic content to populate the area. The aim is for the inpainted area to be both visually and semantically realistic. The applications of these techniques are wide, including removing undesired elements from an image and recovering deteriorated areas.

1.2. SinGAN

SinGAN is a state-of-the-art generative adversarial model developed by Rott Shaham *et al*. [5]. SinGAN can learn from just a single image to generate high-quality images that have the same visual characteristics as the original image. SinGAN uses a series of generative adversarial networks (GANs) which learn the patch distribution at different scales in the image. The originality of SinGAN is that, unlike other single image GANs, it is not restricted to textures and can generate semantic content. Indeed, SinGAN can produce samples of different sizes and aspect ratios with significant variability, whilst retaining the textures and general contextual content of the original image.

Rott Shaham *et al*. [5] have developed a variety of image manipulation tools. The most prominent of these tools is synthetic image generation from a single natural training image. An example of this can be found below.

Figure 1. Synthetic image generation from a single image [5]

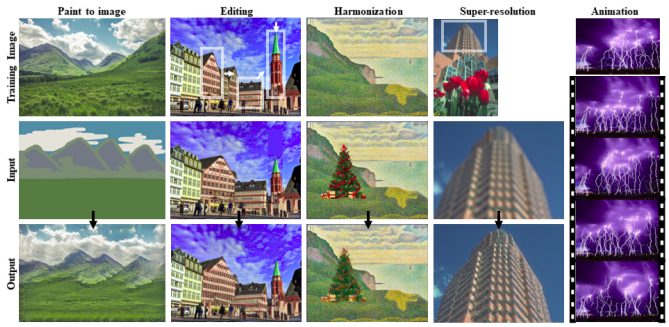

Figure 2. Image manipulation tools [5]

Further manipulation tools include image editing, paint-to-image, harmonisation, super-resolution and the creation of short animations, all from a single natural image. The images in the figure above are an example of the results obtained with these techniques.

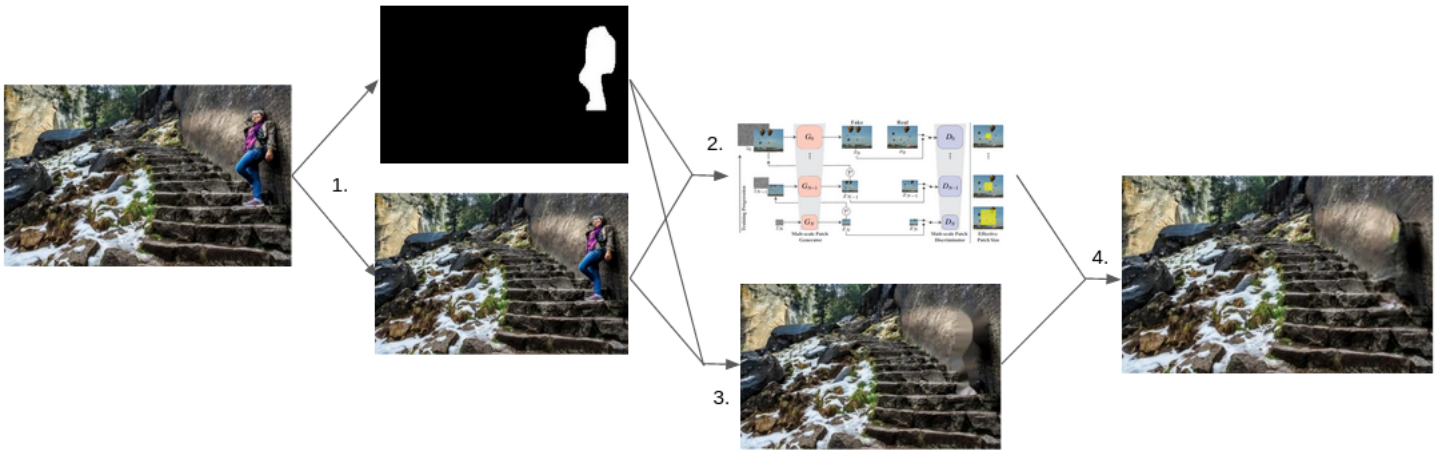

Figure 3. General architecture of our inpainting extension

2. Our method

Based on the image manipulation tools implemented by Rott Shaham *et al*. [5], it is clear that SinGAN has potential to be extended to the challenge of image completion. In this project, an inpainting extension of SinGAN was created. This extension is given a single natural image and a binary mask which identifies the area to be inpainted. If the user's objective is to inpaint a specific category of object which is present in the 91 categories of the COCO 2017 dataset [3], an option can be activated to use an R-CNN model developed by Szegedy *et al*. [2] to segment the object in the initial natural image and to automatically generate the mask. Alternatively, the user can provide his own mask.

Figure 3. illustrates how the inpainting extension functions, the steps of which are detailed below.

1. Create a mask of the area that needs to be inpainted.
2. Train an adapted version of SinGAN on the masked image
3. Create an initial naively inpainted image using Navier-Stokes inpainting [1]
4. Inpaint the naive image using our extension of SinGAN and the trained inpainting-specific model

Inpainting the area with the Navier-Stokes method fills the area with colors that make sense given the surroundings of the masked area. Our inpainting extension then generates colour, texture and content from these initial colours.

The component of SinGAN used by Rott Shaham *et al*. [5] to train the model was adapted to our inpainting objective. The model is only trained on the non-masked part of the image, at all scales. The naive Navier-Stokes inpainting method which initially fills the masked area before it is fed into the proper inpainting extension was imported from Open-CV [1]. The inpainting extension was developed by our team, although inspired by SinGAN's harmonisation tool.

3. Evaluation methodology

In order to evaluate our model, we have chosen to narrow down the challenge of image completion to two specific tasks, namely small defect recovery and the automatic removal of human silhouettes from natural images. The small defect recovery task involves the masking of a 40 by 40 pixel area with plain white colour and its subsequent filling with realistic content. The removal of human silhouettes task involves automatically segmenting the human silhouette and filling in this area with a realistic background. In both case, we have chosen to limit our approach to a single defect and silhouette per image.

In order to quantitatively assess the quality of our inpainting extension, we have used the peak signal-to-noise ratio (PSNR), the structural similarity (SSIM) and the single image Fréchet inception distance (SIFID). The PSNR is an absolute error estimator and measures the quality of reconstruction of a damaged or compressed image and therefore was only used to assess performance in the small defect recovery task. The SSIM is a perceptual error estimator which considers both structural information and perceptual information, such as illumination and contrast. The SIFID is an adaptation of the Fréchet Inception Distance (FID) to the single image setting developed by Rott Shaham *et al*. [5] which evaluates the internal patch statistics relative to those of the real image.

Two existing inpainting solutions were chosen to serve as benchmarks for our inpainting extension. The first, developed by Telea [6] in 2004, is based on the fast marching method and served as a baseline. The second, developed by Nvidia [4], is a state-of-the-art method which uses a CNN with partial convolutions.

4. Quantitative analysis

The results of the quantitative analysis are in the table below.

Type	PSNR (n=8)	SIFID(n=16)	SSIM (n=16)
Telea [6]	33.41(4.20)	$1.1e^{-5}(1.6e^{-5})$	0.96(0.05)
Nvidia [4]	28.96(12.12)	$1.0e^{-5}(1.4e^{-5})$	0.95(0.05)
Ours	31.78(4.33)	$1.0e^{-5}(1.4e^{-5})$	0.95(0.05)

Table 1. Quantitative performance: mean (standard deviation)

According to these statistics, none of the three inpainting tools in our comparison perform significantly better than the others. Indeed, although the mean PSNR, SIFID and SSIM are different for the three inpainting tools, their standard deviations are too high to be able to conclude that the means are statistically different. This has been confirmed by independent sample T-tests at a level of $\alpha = 0.05$. Greatly increasing the sample size would increase the power of the T-tests but may not be sufficient to conclude that there is a significant difference in means. Moreover, it takes 2-3 hours depending on the size of the image to train the model on a new image, consequently increasing the sample size is costly in terms of GPU time.

Furthermore, we will show that in our project the PSNR, SSIM and SIFID metrics were not always able to reliably quantify the perceptual quality of inpainting.

Figure 4. Left: Ours (PSNR: 40.00) Right: Telea (PSNR: 39.45)

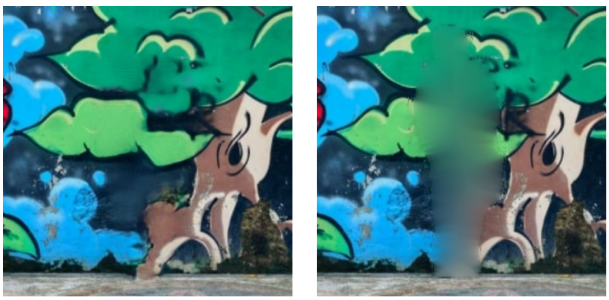

Figure 5. Left: Ours (SSIM: 0.84) Right: Telea (SSIM: 0.86)

Figure 6. Left: Ours (SIFID: $3.92e^{-5}$) Right: Telea (SIFID: $4.84e^{-5}$)

The three pairs of images in this section are comparisons of an image inpainted using an untuned Telea model and with our model. For the three images, the quantitative metrics indicate that the Telea model produced a better result, despite the images from our model being clearly more realistic to the human eye. Although these are only examples and these metrics are known to be useful for a variety of image treatment tasks (such as assessing the precision of a compression tool), in our case, they do not appear to be sufficiently consistent to allow us to draw conclusions on the quality of our inpainting extension versus our competitors.

5. Qualitative analysis

Since the results in the previous part were not reliable, a qualitative evaluation was warranted to assess the quality of our results compared to those of our competitors and how close our inpainted images were to real images. Since the best tool to determine the realism of an image is the human eye, we created a questionnaire divided in two parts. **Link for the questionnaire here.**

In the first part, we wanted to compare our model with that of Nvidia[4]. In each question we showed the respondent an image inpainted by our model and the same image inpainted by Nvidia. The respondent was asked to choose which image was the most realistic. 6 pairs of images were included in this part, 3 consisting of inpainted human silhouettes and 3 consisting in recovered defects.

In the second part, we wanted to compare our model to real images. In each question, we showed the respondent one image, which was either inpainted or real. The respondent was asked whether the image had been modified by our model or whether it was a real image. There were 8 questions in this part, 4 real images and 4 inpainted images.

In order to have fair results, in the first part, we included all the images that we trained our inpainting extension on since the beginning of the project, excluding only those were our extension or Nvidia's produced an absurdly bad result or when a part of the silhouette was missed by the mask. Those used in the second part were also excluded. In

the second part, we included some of our best images and real images from the Web with similar subjects.

In total, 36 people answered our questionnaire. In the first part, respondents found that, over the 6 images, Nvidia provided more realistic results than our inpainting solution (at a 95% confidence level). Indeed, respondents found Nvidia images more realistic 59% of the time. It is clear however that our solution is able to compete with Nvidia for certain types of images, as more respondents found our inpainted images more realistic than Nvidia's for 3 images out of 6. Our solution is comparable to Nvidia's for human silhouette removal where the background is not too complex but struggles to recover defects which cover highly semantic areas.

In the second part, there was only one image were more people thought it was a real image, out of the 4 inpainted images. For each of the 4 real images, more respondents thought it was real. Over all the inpainted images, respondents answered they were real 42% of the time, whereas, over all the real images, respondents answered they were real 72% of the time. The difference here is clearly significant.

These results are very insightful, they show that our model is almost as realistic as that of the state-of-the-art inpainting method developed by Nvidia, delivering the same quality of inpainting for the silhouette removal task. The difference in inpainting performance between both models is certainly due to the differing models employed by the two models, as well as the training process: our model only trains on one image whereas Nvidia's model trains on hundreds of thousands images. This shows how much information can be obtained from only one image but also illustrates its limits, as our model is incapable of hallucinating missing semantic content for large holes. A downside of our model is that it needs to train for 2-3 hours for each image to be inpainted. The second part of the questionnaire shows that some of our best images manage to deceive the human eye. In particular, it must be considered that the respondents were allowed unlimited time to find details showing that the image was modified, but if the model is used for example to hide objects that were not supposed to be present in movie scenes, those modifications could go unnoticed. A selection of images inpainted with both Nvidia's tool and our solution are on the next page.

6. Conclusion

We have created an inpainting extension to SinGAN which from a single natural image and a binary mask is able to generate realistic content to fill the masked area. Conclusions could not be drawn based on the quantitative metrics used to assess the quality of our extension, due to small sample sizes and seemingly unreliable metrics. Our qualitative survey however indicated that our extension is able to satisfactorily inpaint medium-sized holes in natural images, performing at a level similar to that of a state-of-the-art inpainting solution but not being able to consistently fool the respondent into believing that the inpainted images are true. Moreover, the quality of the inpainting depends greatly on the plausible nature of the content to be generated in the masked area: our model struggles to reconstruct unique semantic content not present elsewhere in the image but does particularly well at generating diverse textured and semantic content which are represented elsewhere in the image. Finally, it is clear that a single-image, GAN-based inpainting model suffers from two inherent limitations: it is constrained in terms of the content it can generate due to the limited semantic content it is exposed to and it requires substantial time to train on the single image.

The following table explicitly describes the contribution of the members of the team to the different stages of the project.

	David	Pierre
SinGAN exploration and documentation	50%	50%
Development of the inpainting extension	25%	75%
Development of the human segmentation extension	100%	0%
Production of inpainted images	50%	50%
Production of the survey	75%	25%
Evaluation of performance with metrics	25%	75%
Writing of the report	50%	50%

Table 2. Participation of each student

References

[1] Bertalmio, Marcelo, Andrea L. Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. 2001.

[2] Sergey Ioffe Christian Szegedy and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. 2016.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[4] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions, 2018.

[5] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. *ICCV 2019*, 2019.

[6] Alexandru Telea. An image inpainting technique based on the fast marching method. 2004.

Figure 7. Left: Original Centre: Ours Right: Nvidia

Figure 8. Left: Original Centre: Ours Right: Nvidia

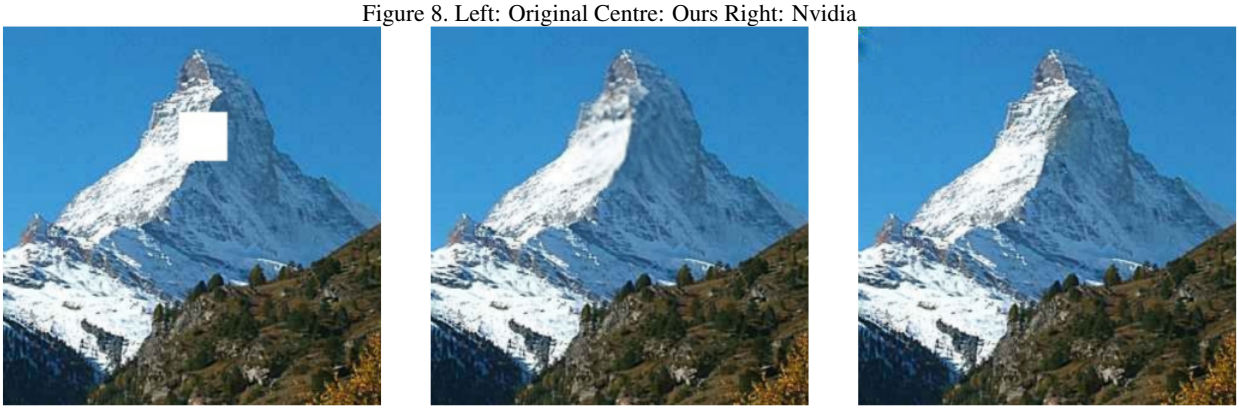

Figure 9. Left: Original Centre: Ours Right: Nvidia

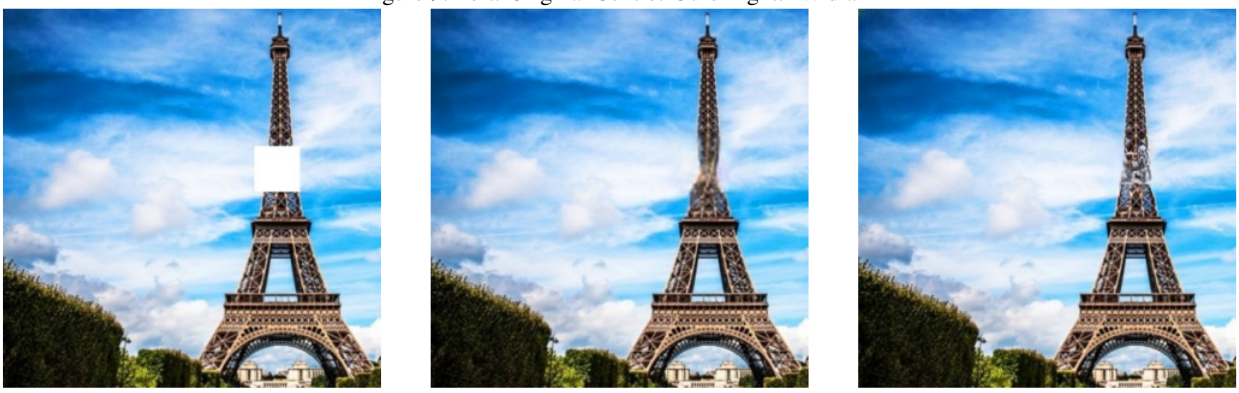

Figure 10. Left: Original Centre: Ours Right: Nvidia