Restoration of Archival Images Using Neural Networks

Raphaela Heil[1], Fredrik Wahlberg[2]

[1]Dept. of Information Technology, Uppsala University, Lägerhyddsvägen 1, 752 37 Uppsala, Sweden
[2]Dept. of Linguistics and Philology, Uppsala University, Thunbergsvägen 3H, 751 26 Uppsala, Sweden

Abstract

Substantial parts of the image material of today's digital archives are of low quality, creating problems for automated processing using machine learning. These quality issues can stem from a multitude of reasons, ranging from damaged originals to the reproduction hardware. Modern machine learning has made automatic "restoration" or "colourization" readily available. Curators and scholars might want to "improve" or "restore" the original's quality to create engagement with the artefacts. However, a fundamental problem of the "restoration" process is that information must always be added to the original, creating reproductions with a synthesized extended realism.

In this paper, we will discuss the nature of the "restoration" or "colourization" process in two parts. Firstly, we will focus on how the restoration algorithms work, discussing the nature of digital imagery and some intrinsic properties of "enhancement". Secondly, we propose a system, based on modern machine learning, that can automatically "improve" the quality of digital reproductions of handwritten medieval manuscripts to allow for large scale computerized analysis. Furthermore, we provide code for the proposed system. Lastly, we end the paper by discussing when and if "restoration" can, and should, be used.

Keywords

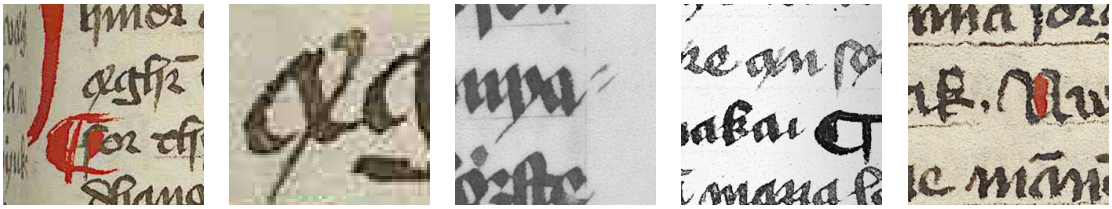
digitization, digital restoration, machine learning, image processing

1. Introduction

As our archives are housing ever-increasing volumes of digital material, the need for automated processing is ever more evident. Automated processing for text search in photographed text [1], scribal attribution [2], and production year estimation [3] are becoming increasingly useful. A common issue posing a significant problem to these endeavours is the quality of some digital reproductions, where old cameras or damaged originals can prove to be impossible obstacles. Digitized historical manuscripts sometimes contain a variety of reproduction deteriorations, such as stains, discolouring and compression artefacts. Some examples of such degradations to the image material are shown in Figure 1. While some types of deteriorations may not impact the work of a trained scholar, they pose a significant challenge regarding legibility and interpretability for both laypersons and computerized processing.

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022

raphaela.heil@it.uu.se (R. Heil); fredrik.wahlberg@lingfil.uu.se (F. Wahlberg)
0000-0002-5010-9149 (R. Heil); 0000-0002-5306-1283 (F. Wahlberg)



(a) (b) (c) (d) (e)

Figure 1: Examples of types of degradations: a) JPEG Compression, b) Detail view of (a) with compression artefacts causing discolourations on the page surrounding the ink stroke, c) Greyscale photography, d) Microform, e) Warping.

Generally, artefacts in digitized manuscripts can be divided into two distinct categories, deteriorations inherent to the material and deteriorations introduced during the digitization process. The former, deteriorations of type I, can often be attributed to the manuscript's age, archival conditions, or handling over time. This type of deterioration can typically be found in all kinds of digitized, historical material. In contrast to these, deteriorations of type II are rooted in the way a manuscript was digitized. Differing, or lack of, image capture protocols can result in issues such as inconsistent illumination or viewpoint, as well as deteriorations caused by the storage process, such as image compression artefacts and colour skewing. Type II is often found in older digitizations, due to technological and procedural constraints. It can however also occur while using modern imaging equipment, such as mobile phones. This classification is not an absolute dichotomy, however, as in the case of some manuscripts on parchment that would need to be flattened for proper photography. This would ruin the original in some cases, where the spine was never designed to allow for being fully opened as the parchment aged.

While the effect of both types of deteriorations can be reduced using classical image processing approaches, these methods are usually highly specialized and require frequent human intervention. They therefore have to be combined into complex pipelines, together with detection of degradation type, and be specifically designed for subgroups of images within the same book. These pipelines seldom generalize well and require both significant on-site expertise and frequent quality inspections. This is both laborious and expensive. One solution for mitigating some of the most common road blocks on the path to automatic processing is specialized cleaning using machine learning [4], which we will focus on in this work.

Another area where "enhancing" digital material is useful is for creating engagement with museum audiences [5]. This can include automatic colourization [6], increasing image resolution [7], or adding video frames [8]. A modern film camera doesn't suffer from the blurriness coming from long exposure time, uneven frame timing, or colour imbalances from chemically developing the film[1]. These limitations can largely be overcome using modern machine learning techniques for video and image enhancement. Some audiences can perceive "enhanced" video as much more engaging, as it looks like something that could have been filmed in the contemporary world they experience. Old battlefields or everyday streets come to life, giving the onlooker a sense of closeness to a lost era.

[1]It does, however, suffer from rolling shutter effects, explaining why some mobile phone videos of helicopters don't show the rotors moving, but that's outside our scope.

A problem for the analysis and "enhancement" described above is that wherever information is missing (colour, video frames, manuscript holes, etc.), it will have to be "guessed" and filled in. This is nothing new to scholars studying any material, as an educated guess can be cultivated though careful study, which is why we'd rather put our trust in a philologist or historian than a layman when it comes to interpreting an incomplete manuscript (or even a complete one). We use the provocative term "guess" here to illustrate the core issue: do you trust the "educated guess" of a computer? This is the fundamental limitation of any computerized "enhancement" technique, as we are forced to put our trust in an opaque process, driven by what often seems like some esoteric mathematical magic. We will argue below that this trust should be a function of the application area and not an automatic choice following from technological positivism. Sometimes adding information to the reproduction of the original creates a "fake", sometimes it improves your analysis. This is also why we choose to put words implying improvement in quotes throughout this paper.

2. Restoration using machine learning

To better explain the process of "restoring" images, and its limitations, we want to start with discussing how images are created and represented in the machine. After this, we can go on to more fully critique the processes and machine learning models of "restoration".

2.1. The nature of a digital image

In the machine, a digital image is represented as coloured areas, called pixels, usually organized in a quadratic grid structure[2]. If you zoom in on any photo, you will usually be able to see this grid structure[3]. A digital camera creates an image by recording the energy from photons (light particles) hitting the image sensor. These photons create a current in the sensor depending on their number and frequency. This process is very similar to the light detection in an eye. Just like in a human eye, there is specialized hardware that detects light of different frequency bands, i.e. colours. The layout of the sensor matches the layout of the pixels in our digital photo.

In the machine, the colours of an image are (usually) represented by intensities in three colour bands for each pixel. These are red, green, and blue. The choice of the colour representation is a trade-off between what a human can see and the technical limitations of computer screens. As there are a lot of colours in the world humans can't see, there is no need for us to fill our camera storage with this information. The colour band intensities are, for historical reasons, most often represented as three integer numbers between 0 and 255. This allows us to, for example, encode a pixel in cornflower blue, with colour intensities [100, 149, 237], as the binary number `011001001001010111101101`. This binary number is what is actually stored in the computer's memory. An illustration of the colour bands is shown in Figure 2, where the three colours have been separated into images for the respective colour bands. The left-most image in the figure is a detail of a cherry blossom, created by putting the three images to the right on top of each other. If you zoom in on the respective colour channel images, you can read (as

[2]While other types of image grids exist, they are very rare.

[3]This statement is not entirely true, as some devices interpolate "intelligently" by add pixels to your photos when zooming in.

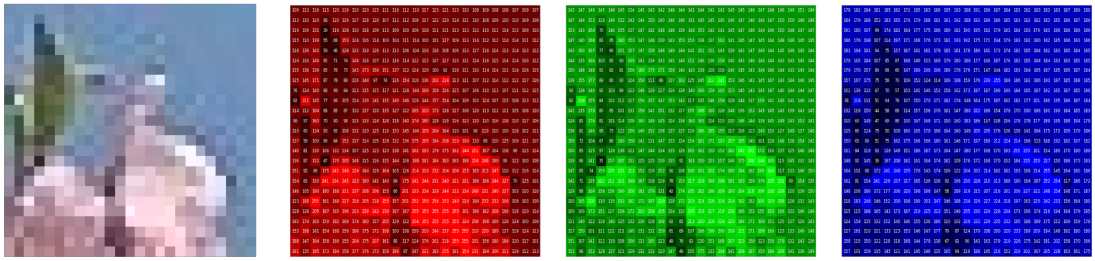


Figure 2: An illustration of colour channels in a digital image, where colour is digitally represented as a mixture of three base colours. The left-most image is a detail from an image of a cherry blossom. The following three images are the Red, Green, and Blue (RGB) colour channels of the same image. If you zoom in on the images to the right, you can see the intensity values, between 0 and 255, for each pixel and colour channel. This, so called RGB encoding, of colour is by far the most common type of colour encoding today.

an illustration) the intensity values for each pixel. Another important point illustrated by the cherry blossom image is that it is often very hard to identify the content of a digital image from a close-up. While avoiding getting too philosophical, it is beneficial to remember that the digital image only serves as an inspiration for the image conjured up in our brains. While humans have this ability for working with images and interpreting the three-dimensional world from a two-dimensional reproduction, the computer has no such evolutionary history or obvious way to learn. Luckily for us, computerized image analysis, a subfield of computer science and mathematics, has been working for decades on building a stack of algorithms which lets us go from the binary ones and zeros in memory, to the image manipulation of modern processing software.

Older camera film only allowed for different greyscale colour schemes. When colourizing such images today, it can be hard to know what the original colour was. As humans, we can often infer the colour from context. We would recognize the Union Jack in a World War 1 image and fill in the colours in our minds. When doing a digital conversion to greyscale from the three RGB colour channels, we mix the colour intensities as $0.299 \cdot r + 0.587 \cdot g + 0.114 \cdot b$, where the letters (r, g, b) are intensity values for their respective colour channel. This gives a greyscale intensity corresponding to the original colour. There are, however, many triplets of r, g, and b that give the same greyscale intensity. In Figure 3, we have taken the Swedish flag (top left) and converted it to greyscale (top right). We then did the same thing with the Scania flag (bottom left), which is a mix of the Swedish and Danish flags. It turned out that the Swedish and the Scania flags look the same in greyscale (top and bottom right). This problem becomes almost impossible to solve when we have less contextual clues, e.g., embroidery on traditional clothing or petals of extinct flowers.

To overcome this problem, a machine learning model for automatic colourization would have to infer colour information from image content. This opens for new challenges, like identifying objects in the image and correlating how they look to some training material. Luckily, this is precisely what has been done.

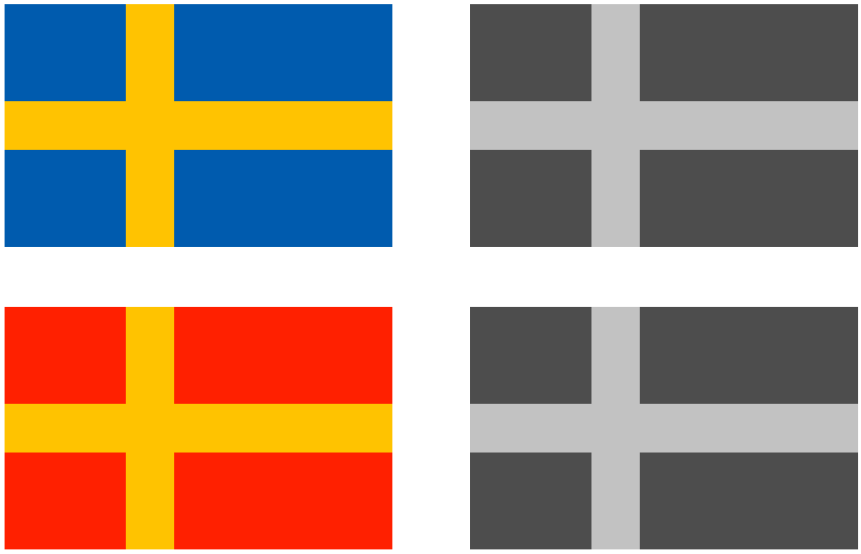


Figure 3: An illustration of how very different colours can look the same in greyscale. The Swedish flag (top left) and the Scania flag (bottom left), a mix between the Danish and Swedish flag, looks the same in greyscale (right).

2.2. Generative machine learning models

In the field of machine learning, the word "model" is an abstract category of algorithms and processes that describe some set of data. In its simplest form, a model can be a spam filter that finds particular words or phrases, and then classifies an email as junk mail. A popular distinction between artificial intelligence (AI) and machine learning (ML) is that the former is a cleverly designed algorithm, while the latter is cleverly designed to *learn* from a set of training data. This means that while an old, AI-based, spam filter would be given specific search terms and phrases, a spam filter built on ML is trained on a set of emails that a human has labelled as junk mail or not. The ML-based spam filter would then try to generalize from training data collected "in the wild". This gives the ML model the advantage of being able to determine its own set of classification rules that are relevant to the task, without having to rely on extensive human efforts. A fundamental characteristic of an ML model is that it must be trained, which can be quite expensive in terms of data and processing power.

Generally, machine learning models are grouped into two major categories, discriminative and generative models. The former deliver categorizations, or classifications, given some data. The initial spam filter example, is a discriminative model, as it does not describe what an email might look like, but only makes a decision whether an email should end up in your junk folder. It can only go in one direction: email input to categorization. This is in contrast to generative models, which, as the name indicates, might be able to generate a full, authentic-looking, email that would be classifiable as either spam or not spam.

Various strategies have been proposed to create generative models, however the approach that is relevant to this paper is *adversarial training*[9]. Generative adversarial networks (GANs)

Figure 4: Four entirely synthetic images of people and animals, generated by a generative adversarial network (StyleGAN2 [11]). Images retrieved from https://thispersondoesnotexist.com/, https://thishorsedoesnotexist.com/ and https://thiscatdoesnotexist.com/, respectively.

are a group of models that aim to generate synthetic data, e.g. images, that would plausibly fit into a collection of real examples, also called the domain of the task. The domain of a spam filter is emails, but the domain where GANs excel is images of natural scenes. Generally, these models consist of two main components, a generator and an adversarial discriminator, that are placed in competition with each other during the learning phase. In this competition, on the one side, the generator tries to create synthetic, albeit plausible data, e.g. images, from the target domain. On the other hand, the discriminator is tasked with identifying whether a presented piece of data is real, i.e. taken from the target domain, or fake, i.e. produced by the generator. Both competitors try to "outsmart" each other by improving their performance on the respective task. In the ideal case, the generator will reach a generative quality at which the discriminator cannot distinguish real from fake. The generator can then be used to create fake images, and is often so good at doing this that humans are fooled. Note that the training data doesn't need to be strongly curated, one only needs to make sure that the GAN has enough images from the intended domain. As an illustration of this, Figure 4 shows some synthetic images generated with the methodology above. All of the depicted images of people and animals were generated using a model called StyleGAN2 [10, 11].

The outlined general idea of GANs can be put into a real-world context by considering the analogy of an art student and a critic, where the former takes the role of the generator, creating paintings, while the latter represents the discriminator, determining whether an artefact is worthy of a spot in the town's art gallery[4]. At the start of the training phase (in our analogy), both our student and critic start out by being completely unskilled at their respective crafts. Hence, the critic's job is fairly easy in the beginning. To make this analogy work, our critic must be excellent at giving feedback to the student on how they can improve. The critic will now pick up paintings randomly from either some collection of acceptable art, the domain, or some paintings made by the student. After careful deliberation, our critic will make a decision on if this is acceptable art. If the randomly picked painting was made by our student, the critic will give them feedback on how to create art that looks more like the paintings in the domain. This process will then continue until our critic, who is constantly getting better at their work, can no longer distinguish the domain paintings from at artwork by our student. If this was more than an analogy, an objection to this setup might be that the student isn't really encouraged to

[4]Goodfellow et al. describe these roles as counterfeiters and police, respectively [9]

be creative, i.e. actually create art. This is true for the machine learning model too. The model doesn't learn creativity, it learns to imitate.

It should be noted that machine learning is an applied field in the sense that models are researched and trained in order to solve some task. It is widely accepted that "all models are wrong, but some are useful" [12]. When a model learns to imitate some training data, the type of data that is imitated is of crucial importance to "enhancement" using the adversarial generative techniques described above. Let's say we have a photo, taken in the 1930s. It is blurry and in greyscale. One way of "enhancing" the quality of this photo is to make the generative model find an image, with both high resolution and colour, that matches our original. How do we determine the best match? We can think of this as the model generating images, we convert them to blurry greyscale, and then compare, pixel by pixel, to the original. Note that some operations that are very time-consuming for a human can be well adapted to a computer's capabilities. After some searching we could always find multiple images that, after conversion, would match the original. This is unavoidable, as the original simply doesn't contain the information needed to "enhance" detail, e.g., if there is a person in the background or just a trick of the light. Though the "enhanced" photograph can be very convincing, as we show in Figure 4, this should not be taken as evidence for that it shows something that is not conjured up by the model, like in some American crime dramas. However, if the model has been trained on images very much like the original, perhaps through high quality modern re-enactment or some timeless physical phenomenon, it is likely to synthesize something highly plausible.

3. Cycle-consistent Generative Adversarial Networks (CycleGANs)

In order to "restore" degraded archival manuscript pages, we propose to employ a type of GAN, commonly referred to as cycle-consistent generative adversarial networks (CycleGANs) [13]. These models have found prior applications in a variety of image translation tasks, such as the presentation of photos in the style of famous classical painters, e.g. Van Gogh [13], the transformation of images like portraits and animal photos to traditional Japanese flower arrangements [14] and the removal of strike-through artefacts from handwritten words [15]. Additionally, CycleGANs have been used to remove certain degradations, such as stains and watermarks, from printed documents [16].

The general structure of a CycleGAN is illustrated in Figure 5, using the task of image restoration as an example. Concretely, this approach consists of two regular GANs that are trained in conjunction with each other. As shown in the illustration, each of the two generators, named 'restore' and 'degrade', are concerned with generating restored, respectively degraded, images. Discriminator A assesses whether a given clean images is a genuine high-quality one or was created by the generator, while discriminator B determines whether a presented degradation is genuine or generated.

To demonstrate the flow of images through the system, we trace the path of a degraded image in the following (Figure 5, left). Initially, the degraded image is fed into the generator 'restore', which uses it as a basis to produce a restored image. This image is then assessed by the discriminator A. In addition to this, it is processed by generator 'degrade', which returns it

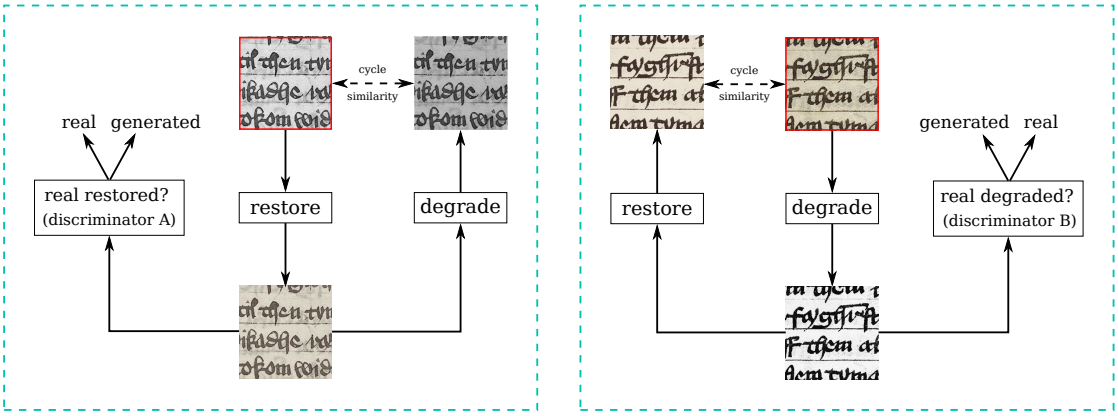


Figure 5: General structure of a CycleGAN. Dataflow is being demonstrated via an example of image "restoration" (left) and degradation (right). Starting points of each cycle, i.e. input images, are marked with a red border, arrows indicate input and output of data, and 'restore' and 'degrade' represent the two generators, performing the respective transformations.

to a degraded state. Besides the discriminator's assessment, feedback regarding the generation quality is provided to 'restore' by comparing the original input image with the restored and subsequently degraded image. The latter comparison aims for a high similarity, which is referred to as cycle-consistency, hence the name of the model.

In parallel to this pass, a second cycle takes place, starting from a clean image. This process is shown on the right in Figure 5, following the clean input image through the generator 'degrade' to a degraded output and back through generator 'restore' to a restored image, to close the cycle via a comparison of the latter with the input. The same feedback mechanisms as in the first cycle are applied to the respective generator, using the respective discriminator.

Various implementation for the generators and discriminators exist and are often, to some degree, determined by the task at hand. In this work, we employ dense U-Nets [17], via the implementation from [18] for the generators and the traditional discriminators, proposed by [13].

3.1. Data

In order to train and evaluate the proposed approach, we randomly selected 17 manuscripts from *Manuscripta*[5], a digital collection of medieval and early modern manuscripts, provided by the *National Library of Sweden*. The selected manuscripts are dated between ca. 1300 and ca. 1526 and contain texts in Old Swedish. A total of 150 pages were randomly selected from these manuscripts and split into 75 pages for the training, 25 for the validation and 50 for the test set. While preparing these splits, we ensured that images from one manuscript would only be present in exactly one split, in order to avoid information leakage.

With the aim of closely representing real archival image conditions, we consider the following artificial degradations (cf. Figure 1) in our experiments:

[5] https://www.manuscripta.se/

JPEG Compression Page images are converted to a JPEG representation with varying compression levels, randomly sampled from the range [65,85].

Greyscale The standard, weighted greyscale transformation is applied to the input image.

Microform Firstly, the given image is converted to greyscale, following the same weighted approach as above. Subsequently, the contrast is adjusted via a sigmoid correction [19]. Lastly, the illumination of the microform during digitisation, is simulated by superimposing a mask which darkens the image towards the edges.

Warping A small elastic transformation [20], with alpha and sigma randomly drawn from [15,25], respectively [4,6] is applied, resulting in warping and slight displacements of the image content.

Following the taxonomy of degradations, outlined in the introduction, *Warping* can be categorised as type I, while the other three are examples of type II. Regarding the altered images, it should be noted that *JPEG Compression* and *Warping* produce RGB images, while *Greyscale* and *Microform* result in single-channel representations. For ease of training, outputs from the latter two transformations are repeated three times and stacked, to result in a three-channel format. Implementations for *JPEG Compression*, *Greyscale* and *Warping* were provided by [21], while *Microform* is a custom implementation. The code can be found in the accompanying repository (cf. Appendix A).

For the preparation of the validation set, each page was altered individually by each of the four outlined approaches. Subsequently, one patch random patch of size 256 by 256 pixels was cropped from each page, for each of the four augmentations. This results in a set of 100 image patches, each of which are stored together with the corresponding patch from the unaltered page, for comparison.

The test set was prepared in a similar fashion, however instead of a single patch per page and augmentation, ten random, disjoint, i.e. not overlapping, patches were selected, resulting in a total of 2000 patches.

3.2. Neural Network Training Protocol

We train the outlined CycleGAN for a total of 60 epochs, using the Adam [22] optimizer with a learning rate of 0.001. Each epoch entails the sampling with replacement of 300 pages from the training set. 50 percent of these are artificially deteriorated, using a randomly selected approach from the ones outlined above, while the other 50 percent are left unaltered. One randomly located, square patch of width 256px is cropped from each of the 300 pages. Deteriorated patches are input into generator 'restore' and discriminator B, while clean ones are supplied to generator 'degrade' and discriminator A. Following each training epoch, we assess the restoration performance on images from the validation set, via the Root-Mean-Square Error (RMSE) and the Structural Similarity Index Measure (SSIM)[23]. The model checkpoint exhibiting the best validation performance is retained and used for evaluation on the test set.

Degradation	Degraded SSIM	Restored SSIM	Degraded RMSE	Restored RMSE
Microform	0.8046	0.8301	0.3717	0.2849
Greyscale	0.9909	0.9369	0.0921	0.1033
Warping	0.8928	0.8502	0.0702	0.1144
JPEG	0.9484	0.9099	0.0353	0.0857
Overall	0.9092	0.8818	0.1423	0.1471

Table 1
Quantitative results for the "restoration" task. Both SSIM and RMSE range between zero and one. For the former, higher values are better, while for the latter, values closer to zero are.

3.3. Evaluation

In order to evaluate the "restoration" performance of our proposed approach, the chosen model checkpoint was applied to all patches in the test set. A hand-picked selection of model outputs, one per degradation type, is shown in Figure 6 (bottom row), together with the original state of the patch (top row) and the altered version that was provided as input to the 'restore' generator (middle row).

As can be seen from the samples, the generator successfully transforms the *microform* and *greyscale* patches into coloured images. The range of displayed colours is slightly diminished, one could say *muted*, as compared to the original images, however none of the patches display extreme or unexpected colours. For the warped patch, a de-warping effect is not immediately apparent, however it can be noted that the generator has adapted the colour slightly. A similar colour effect can be observed for the JPEG-compressed patch. In contrast to the warped one, however, a "restoration" effect, in the form of smoothing, and reduction of blocking and ringing artefacts, is visible. Overall, the above observations hold for the majority of the test patches of types *microform*, *greyscale* and *warping*. Results for the *JPEG* compression are more diverse in quality, but some level of improvement or smoothing is generally observable.

To provide a more comprehensive overview over the model's "restoration" performance, Figure 7 illustrates two samples for which the generator provides a convincing "reconstruction" of the background colour but fails to correctly represent the use of red ink, visible in the original patch. These examples tangibly demonstrate how the extent of the training data can influence the model's "restoration" performance. As red ink is used sparingly, only to highlight selected words or phrases, it is not represented as frequently in the training data as regular, black ink. The model will therefore exhibit a strong tendency to colour darker areas, generally corresponding to some form of ink in the *microform* and *greyscale* images, in black or darker greys instead of other potential ink colours. Training the model on a dataset with more diverse shades of ink would potentially mitigate this issue.

Besides the qualitative evaluation, we also present a brief quantitative analysis below. Table 1 shows the *RMSE* and *SSIM* values for the altered and the "restored" patches, each calculated with respect to the ground truth. Notably, the measures only improve for "restored" patches in the case of *microform* degradations. In all other cases, the performance appears to drop by one to five percentage points. Considering that a large portion of the qualitative results

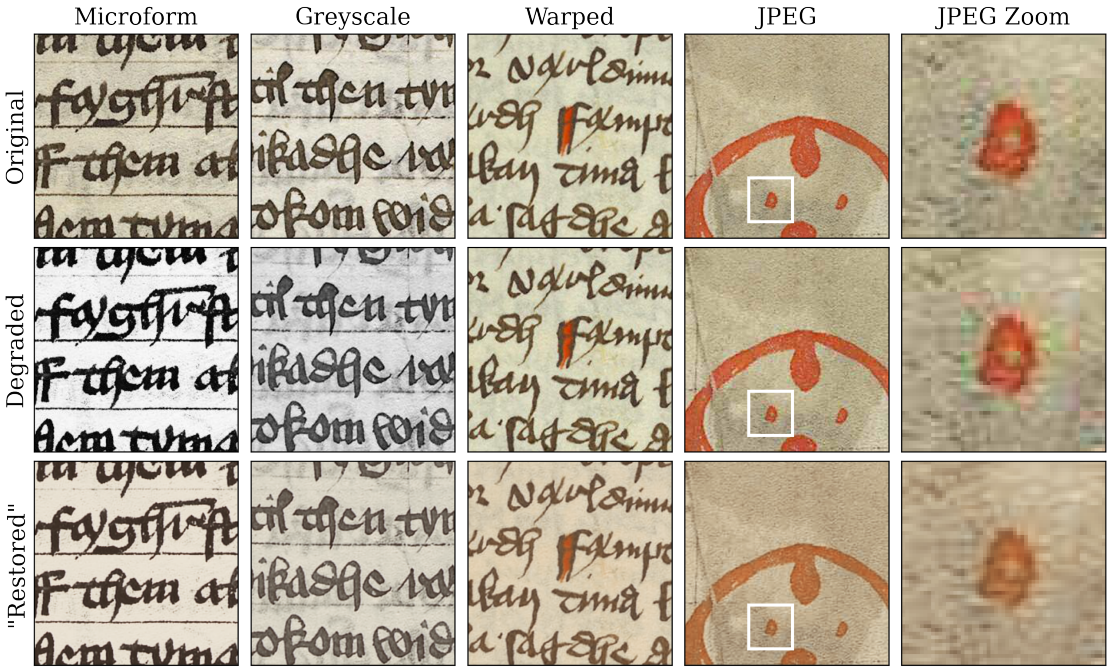


Figure 6: Hand-picked "restoration" samples, demonstrating the performance of the model. The first row shows the original image patch, the second the degraded version that is input into the generator 'restore' and the last row show the generator's output. The last column ("JPEG Zoom") shows a close-up of the box, marked in white in the previous column ("JPEG").

are perceived as reasonable restorations, this raises the question of how the qualitative and quantitative results can be consolidated and put into context with each other.

A major aspect to consider, when it comes to the evaluated metrics, is that these require a concrete reference image, or ground truth, to compare with. Herein, however, lies the crux of the restoration problem: the true original data is not available. One can only attempt to obtain an approximation of the *truth* by making "educated guesses" based on existing clean and artificially altered data. It can therefore be argued that, while these metrics can be useful to get a general idea of the similarity and thereby of a certain degree of quality in the context of artificial data, they should not be considered in isolation. Instead, they should be combined with the qualitative results and examined by experienced scholars in an appropriate context.

4. Discussion & Conclusion

With the continued increase in computational power and capabilities of machine learning approaches, many opportunities arise in the humanities. These methods can not only facilitate novel research questions, but also provide opportunities for engaging with the public. As demonstrated above, machine learning (ML) can act as a convenient tool, but should not be used carelessly or trusted blindly. In order to "restore" information, any model will have to make an *educated guess*, based on the data it is presented during training. Through this, the

Figure 7: Hand-picked "restoration" samples, demonstrating cases where the model proposes a "restoration" that differs substantially from the expected colourization.

"restorations" run the risk of becoming copies of the training data, instead of approximations of the true original form, which we are aiming for in the case of image enhancements.

The incorporation of undesired pieces of information in a "restoration" can have a considerable detrimental effect when research questions should be answered using the processed data. A wrong representation or interpretation will affect the results and conclusions. This effect also has to be considered when data is being refined for consumption by laypersons, for example in the form of an exhibition or teaching material. It is crucial to ensure that the "reconstructions" allow observers to get an appropriate idea of the information and its implications. The initial example of the colour "restoration" of the greyscale flag (cf. Figure 3) serves as a cautionary tale here: whether the Swedish or the Scanian flag is presented to an audience, whatever level of expertise they may hold, could considerably affect their interpretation of the context and any conclusions they draw from it.

As argued above, generative modelling in machine learning can create digital images of impressing fake realities. This is done through a process of learning to imitate a domain of real images, e.g., human faces or landscapes. For museums, "enhancing" historical photographs or film has the potential to create engagement with older image material. As with any generative model, the frames that are filled in, or the pixels that are added, take their inspiration from both the original and the image material in the training set of the machine learning method. Hence, the historical material is merged with modern material to create a synthetic high-quality image. As such, small details in a historical photograph will be "filled in" by data from the training images in its high-quality counterpart. How can we be confident that the computer made the correct choice when filling in this information? If the training data is modern, a square object in someone's hand might get interpreted as a mobile phone, whereas originally, it may have been a cigarette case. When a model is used for sensitive material, the information that it is trained on therefore needs to be curated carefully.

Generative modelling for image "enhancement" is likely to get better and less expensive in the coming years. Making historical film look like it was taken with modern equipment, like

Peter Jackson did with World War 1 footage in "They shall not grow old", is likely to become ubiquitous in museums. As computer scientists, we want to caution against using such methods uncritically. The temptation of technological positivism is strong, with a substantial hype around "artificial intelligence" (what we here call machine learning). It's always warranted to ask about which training data we are "enhancing" from. What do they include, and more importantly, what type of images or people are not included in the training data? The choice of using machine learning "enhancement", and which data to base it on, must follow from the intended application area. If it improves text recognition, "enhancement" can be used fairly safely, as the risk of accidentally adding plausible words is very low. If you are going to improve historical photography, however, make sure the model was trained on similar material to yours. If you want to make out details that weren't really visible in the original, you are not working with the original any more, and your results will be based on a synthetic reality. This last case is the most important, as there are a lot of imaging applications where super-resolution is viable. This does not apply to historical imagery, as the material is much more diverse than CT scans or cartoons. As is often the case with new technology, the potential is great and intriguing. However, in a world of believable synthetics and deepfakes, the need for a trained human eye and careful curation has never been greater.

Acknowledgments

The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

This work is partially supported by Riksbankens Jubileumsfond (RJ) under the project "New Eyes on Sweden's Medieval Scribes", Dnr NHS14-2068:1. A special thanks to the PI, professor Lasse Mårtensson.

References

[1] T. Wilkinson, J. Lindstrom, A. Brun, Neural ctrl-f: Segmentation-free query-by-string word spotting in handwritten manuscript collections, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[2] A. Brink, J. Smit, M. Bulacu, L. Schomaker, Writer identification using directional ink-trace width measurements, Pattern Recognition 45 (2012) 162–171. URL: https://www.sciencedirect.com/science/article/pii/S0031320311002810. doi:https://doi.org/10.1016/j.patcog.2011.07.005.

[3] S. Boldsen, F. Wahlberg, Survey and reproduction of computational approaches to dating of historical texts, in: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), 2021, pp. 145–156. URL: https://aclanthology.org/2021.nodalida-main.15.

[4] R. Heil, E. Vats, A. Hast, Paired image to image translation for strikethrough removal from handwritten words, 2022. arXiv:2201.09633, Under review at DAS 2022.

[5] H. Sommer, Assessing millennial engagement in museum spaces, in: Theory and Practice 1, The Museum Scholar, 2018. URL: http://articles.themuseumscholar.org/tp_vol1sommer.
[6] G. Larsson, M. Maire, G. Shakhnarovich, Learning representations for automatic colorization, 2017. arXiv:1603.06668.
[7] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, CoRR abs/1501.00092 (2015). URL: http://arxiv.org/abs/1501.00092. arXiv:1501.00092.
[8] S. Niklaus, F. Liu, Context-aware synthesis for video frame interpolation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014. arXiv:1406.2661.
[10] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, 2019. arXiv:1812.04948.
[11] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, 2020. arXiv:1912.04958.
[12] G. E. P. Box, Science and statistics, Journal of the American Statistical Association 71 (1976) 791–799. doi:10.1080/01621459.1976.10480949.
[13] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
[14] C. H. Mai, R. Nakatsu, N. Tosa, Developing japanese ikebana as a digital painting tool via ai, in: N. J. Nunes, L. Ma, M. Wang, N. Correia, Z. Pan (Eds.), Entertainment Computing – ICEC 2020, Springer International Publishing, Cham, 2020, pp. 297–307.
[15] R. Heil, E. Vats, A. Hast, Strikethrough removal from handwritten words using cyclegans, in: J. Lladós, D. Lopresti, S. Uchida (Eds.), Document Analysis and Recognition – ICDAR 2021, Springer International Publishing, Cham, 2021, pp. 572–586.
[16] M. Sharma, A. Verma, L. Vig, Learning to clean: A gan perspective, in: G. Carneiro, S. You (Eds.), Computer Vision – ACCV 2018 Workshops, Springer International Publishing, Cham, 2019, pp. 174–185.
[17] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1175–1183. doi:10.1109/CVPRW.2017.156.
[18] N. Pielawski, OctoPyTorch: Segmentation Neural Networks, 2021. URL: https://github.com/npielawski/octopytorch.
[19] G. J. Braun, M. D. Fairchild, Image lightness rescaling using sigmoidal contrast enhancement functions, Journal of Electronic Imaging 8 (1999) 380–393.
[20] P. Simard, D. Steinkraus, J. Platt, Best practices for convolutional neural networks applied to visual document analysis, in: ICDAR, 2003. doi:10.1109/ICDAR.2003.1227801.
[21] A. B. Jung, K. Wada, J. Crall, S. Tanaka, J. Graving, C. Reinders, S. Yadav, J. Banerjee, G. Vecsei, A. Kraft, Z. Rui, J. Borovec, C. Vallentin, S. Zhydenko, K. Pfeiffer, B. Cook, I. Fernández, F.-M. De Rainville, C.-H. Weng, A. Ayala-Acevedo, R. Meudec, M. Laporte, et al., imgaug, https://github.com/aleju/imgaug, 2020. Online; accessed 14-Feb-2022.
[22] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun

(Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6980.

[23] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612. doi:10.1109/TIP.2003.819861.

A. Online Resources

The code used to train and evaluate the proposed CycleGAN is available here: https://zenodo.org/record/6592707.